The American Journal of Human Genetics

Vol. 7 No. 3 *September 1955*

Chances of Excluding Paternity by the Rh Blood Groups[1]

WILLIAM C. BOYD

School of Medicine, Boston University

WHEN making use of blood tests in an attempt to exonerate a supposedly falsely accused man of a charge of paternity, it is helpful to have an estimate of the chances of making an exclusion with a given blood group system, especially the probabilities when the blood group of the accused man is known. A knowledge of such probabilities will influence the number of blood group systems employed, and enable the expense of any given set of tests to be weighed against the probable benefit.

General formulas which enable such calculations to be made for populations possessing any given set of gene frequencies have been derived for the ABO, MN, and MNS systems (Wiener, Lederer and Polayes 1930; Zarnik 1930; Wiener 1952; Boyd 1955). They are of course particularly simple for a system consisting of a simple pair of genes, one dominant (Wiener, Lederer and Polayes 1930; Cotterman 1951; Race and Sanger 1950, 1954). Thus far no such formula has been derived for the more complicated Rh system, and the only information available is a table constructed in 1944 by R. A. Fisher, based on approximate gene frequencies which might apply to the English population, and printed, without explanation of how it was constructed, by Race and Sanger (1950, 1954). These results would obviously not apply to populations with different Rh gene frequencies, and it might be wondered how exact the results, based on approximate gene frequencies, are for the English population. It is the purpose of the present communication to supply the required general formulas.

The first requirement in such calculations is a table showing the frequency with which children of the various phenotypes are born to women of the various phenotypes, when such women are supposed to be mated to men drawn at random from the general population. Such frequencies are given in table 1, which was constructed by examining the children which would be born to women of each phenotype when they receive the various Rh genes in numbers proportional to the frequencies of these genes in the population. (Table 1 would of course apply equally to father-child combinations.) For the sake of brevity, the frequencies of the genes R_2, R_0, R'', r, R_z, R_1, and R' are represented by the letters t, u, v, w, x, y, z in that order (the order given by Fisher 1946). Using these symbols, the expected frequency of the various phenotypes are

Received March 25, 1955.

[1] The research reported in this paper was made possible in part by the use of equipment purchased by Boston University under Contract No. Nonr-492(01) with the Office of Naval Research.

TABLE 1.—FREQUENCY OF CHILDREN OF VARIOUS PHENOTYPES BORN TO VARIOUS MOTHERS

Phenotype of Children	Phenotype of Mother										
	1 cde rh	2 ceE Rh″	3 cDe Rh$_0$	4 cDE Rh$_2$	5 Cde/c Rh′rh	6 CdE/c Rh′Rh″	7 CDe/c Rh$_1$rh	8 CDE/c Rh$_1$Rh$_2$	9 Cde/ C Rh′- Rh′	11 CDe/C Rh$_1$Rh$_1$	12 CDE/C Rh$_1$Rh$_z$
1. cde rh	w^2	vw^2	uw^2	tw^2	w^2z	—	w^2y	xw^2	—	—	—
2. cdE Rh″	vw^2	$v[(v+w)^2+vw]$	uvw	$v[(t+u)(v+w)+tw]$	vwz	$vz(v+w)$	vwy	$v[(v+w)(x+y)+wx]$	—	—	—
3. cDe Rh$_0$	uw^2	uvw	$u[(u+w)^2+uw]$	$u[(t+v)(u+w)+tw]$	uwz	—	$u[(u+w)(y+z)+wy]$	$ux(u+2w)$	—	—	—
4. cDE Rh$_2$	tw^2	$v[(t+u)(v+w)+tw]$	$u[(t+v)(u+w)+tw]$	$tk^2+u(t+v)^2+v(t+u)^2+t^2w$	twz	$vz(t+u)$	$twy+u(t+v)(y+z)$	$tkp+v(t+u)(x+y)+x[u(t+v)+tw]$	—	—	—
5. Cde/c Rh′rh	w^2z	vwz	uwz	twz	$wz(w+z)$	vwz	$wz(u+y)$	$wz(t+x)$	wz^2	wyz	wxz
6. CdE/c Rh′Rh″	—	$vz(v+w)$	—	$vz(t+u)$	vwz	$vz(v+z)$	uvz	$vz(t+x+y)$	vz^2	vyz	vxz
7. CDe/c Rh$_1$rh	w^2y	vwy	$u[(u+w)(y+z)+wy]$	$twy+u(t+v)(y+z)$	$wz(u+y)$	uvz	$u(y+z)^2+y(u+w)^2+u^2z+y^2w$	$y(t+v)(u+w)+x[u(y+z)+wy]+tuz$	uz^2	$y[(y+z)(u+w)+uz]$	$x[u(y+z)+wy]$
8. CDE/c Rh$_1$Rh$_2$	w^2x	$v[(v+w)(x+y)+wx]$	$ux(u+2w)$	$tkp+v(t+u)(x+y)+x[u(t+v)+tw]$	$wz(t+x)$	$vz(t+x+y)$	$y(t+v)(u+w)+x[u(y+z)+wy]+tuz$	$tp^2+xk^2+y(t+v)^2+v(x+y)^2+x^2(u+w)+t^2z$	tz^2	$y[(y+z)(t+v)+tz]$	$x[kp+y(t+v)+tz]$
9. Cde/C Rh′Rh′	—	—	—	—	wz^2	vz^2	uz^2	tz^2	z^3	yz^2	xz^2
11. CDe/C Rh$_1$Rh$_1$	—	—	—	—	wyz	vyz	$y[(u+w)(y+z)+uz]$	$y[(t+v)(y+z)+tz]$	yz^2	$y[(y+z)^2+yz]$	$xy(y+2z)$
12. CDE/C Rh$_1$Rh$_z$	—	—	—	—	wxz	vxz	$x[y(u+w)+uz]$	$x[kp+y(t+v)+tz]$	xz^2	$xy(y+2z)$	$x[p^2+x(y+z)]$

1. $cde = w^2$
2. $cdE = v^2 + 2vw$
3. $cDe = u^2 + 2uw$
4. $cDE = t^2 + 2tv + 2tu + 2tw + 2uv$
5. $Cde/c = 2wz$
6. $CdE/c = 2vz$
7. $CDe/c = 2(uy + wy + uz)$
8. $CDE/c = 2(tx + ty + tz + vx + vy + ux + wx)$
9. $Cde/C = z^2$
10. $CdE/C =$ usually absent
11. $CDe/C = y^2 + 2yz$
12. $CDE/C = x^2 + 2xy + 2xz$

In table 1 the letter k stands for $t + u + v + w$, and p stands for $x + y + z$. The rare gene R_y, although probably present, is assumed to be absent, and in any case it is so rare as to contribute very little to the chances of exclusion. Therefore the very rare phenotype CdE/C (R_yR'), which would be number 10 in a complete listing of the 12 phenotypes distinguishable with 4 sera (Boyd 1954b) is not included. Omitting the non-contributive (for our purposes) R_y from consideration results in considerable simplification of the tables and calculations.

From table 1, knowing the frequencies of the various genes in the population, a table showing the numerical values of the frequency with which children of the various types are born to various women can be constructed. This has been done in table 2, which shows the values for a population having the gene frequencies derived by Fisher (1946, 1947) form the data of Fisher and Race (1946). These frequencies are, to four decimal places,

t	0.1280
u	0.0305
v	0.0170
w	0.3790
x	0.0013
y	0.4361
z	0.0081

The dashes in table 2 indicate that no child of the type indicated at the left could be born to a mother of the type indicated above. In the case of row 1, column 6, and row 6, columns 1 and 3, this is the result of ignoring the gene R_y. This gene is so rare that even if we were to take account of it nothing significant would be contributed to these entries in the table.

Following the generally accepted practice of carrying in computation more decimals than it is desired to retain at the end (Boyd 1954a), all the calculations in this paper have been carried out using six places of decimals, finally rounding off to four. This explains occasional differences in the fourth place between values given in tables 2 and 4 and values which would result from only four-figure accuracy.

It is next necessary to know which combinations of mother and child will exclude

TABLE 2.—NUMERICAL FREQUENCIES OF VARIOUS MOTHER-CHILD COMBINATIONS

Phenotype of children	Phenotype of Mother										
	1 cde	2 cdE	3 cDe	4 cDE	5 Cde/c	6 CdE/c	7 CDe/c	8 CDE/c	9 Cde C	11 CDe/C	12 CDE/C
1. cde	.0544	.0024	.0044	.0184	.0012	—	.0626	.0002	—	—	—
2. cdE	.0024	.0028	.0002	.0019	.0001	.0001	.0028	.0029	—	—	—
3. cDe	.0044	.0002	.0055	.0033	.0001	—	.0106	.0000	—	—	—
4. cDE	.0184	.0019	.0033	.0466	.0004	.0000	.0231	.0329	—	—	—
5. Cde/c	.0012	.0001	.0001	.0004	.0012	.0001	.0014	.0004	.0000	.0013	.0000
6. CdE/c	—	.0001	—	.0000	.0001	.0000	.0000	.0001	.0000	.0001	.0000
7. CDe/c	.0626	.0028	.0106	.0231	.0014	.0000	.1513	.0262	.0000	.0794	.0002
8. CDE/c	.0002	.0029	.0000	.0329	.0004	.0001	.0262	.0383	.0000	.0285	.0004
9. Cde/C	—	—	—	—	.0000	.0000	.0000	.0000	.0000	.0000	.0000
11. CDe/C	—	—	—	—	.0013	.0001	.0794	.0285	.0000	.0876	.0003
12. CDE/C	—	—	—	—	.0000	.0000	.0002	.0004	.0000	.0003	.0003

paternity for a man of given type, allowing for the fact that to mothers of certain types children of some phenotypes will not be born. This information is presented compactly in table 3, by the device of representing the various phenotypes by numbers. The numbers which stand for the various phenotypes are shown along the side of the table, and also at the top.

Note that phenotype 8 is not listed at the top, as a man of such phenotype can not establish non-paternity by the Rh blood groups. Also note that the omission of R_y from consideration means that certain (very rare) exclusions are omitted. For instance if R_y were present, phenotype 6 (CdE/c = Rh′Rh″) would include genotype R_yr (CdE/cde) and such a woman, mated (for example) to a man of type 1 (cde — rh), could produce a child of phenotype 6, which, in the case of men of phenotype 9, 11 or 12 (Cde/C, CDe/C or CDE/C) would exclude paternity. Such possible (but very rare) children have not been included in table 3. On the other hand, since R_y is probably present, although very rare, in our population, no exclusions which could be based upon its absence have been included. For example, if R_y were not present, a woman of type 6 and a man of type 7 could not have a child of type 12, but this and other such exclusions have not been included in table 3.

Table 3 was worked out with the aid of the CDE notation, which is much more convenient than Wiener's for this purpose, but for the benefit of readers who may be more familiar with Wiener's notation, the designations of the phenotypes are given in this system as well.

In the case of the simpler blood group systems, it is possible to derive general formulas for the probabilities of exclusion by noting the phenotypes of children which would exclude a man of one phenotype when accused by a woman of each of the types, and adding all the expressions for their frequency. Thereby one obtains a formula which, since it gives the total probability, for each type of man, of finding mother-child combinations which exclude paternity, gives his chances of establishing non-paternity. Then by multiplying each of these expressions by the general formula for the frequency of the phenotype of the man, one obtains a general formula for the chances of a man, blood group unknown, establishing non-paternity.

TABLE 3.—PHENOTYPES OF CHILDREN ESTABLISHING NON-PATERNITY IN VARIOUS COMBINATIONS OF MOTHERS AND ALLEGED FATHERS

Phenotype of Mother	Phenotype of Alleged Father									
	1 rh cde	2 Rh″ cdE	3 Rh$_0$ cDe	4 Rh$_2$ cDE	5 Rh′rh Cde/c	6 Rh′Rh″ CdE/c	7 Rh$_1$rh CDe/c	9 Rh′Rh′ Cde/C	11 Rh$_1$Rh$_1$ CDe/C	12 Rh$_1$Rh$_2$ CDE/C
1. rh cde	2, 3, 4, 5, 7, 8	3, 4, 5, 7, 8	2, 4, 5, 7, 8	5, 7, 8	2, 3, 4, 7, 8	3, 4, 7, 8	2, 4, 8	1, 2, 3, 4, 7, 8	1, 2, 3, 4, 8	1, 2, 3, 4
2. Rh″ cdE	3, 4, 5, 6, 7, 8	3, 4, 5, 6, 7, 8	5, 6, 7, 8	5, 6, 7, 8	3, 4, 7, 8	3, 4, 7, 8	—	1, 2, 3, 4, 7, 8	1, 2, 3, 4	1, 2, 3, 4
3. Rh$_0$ cDe	2, 4, 5, 7, 8	5, 7, 8	2, 4, 5, 7, 8	5, 7, 8	2, 4, 8	—	2, 4, 8	1, 2, 3, 4, 8	1, 2, 3, 4, 8	1, 2, 3, 4
4. Rh$_2$ cDE	5, 6, 7, 8	5, 6, 7, 8	5, 6, 7, 8	5, 6, 7, 8	—	—	—	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4
5. Rh′rh Cde/c	2, 3, 4, 6, 7, 8, 9, 11, 12	3, 4, 7, 8, 9, 11, 12	2, 4, 6, 8, 9, 11, 12	9, 11, 12	2, 3, 4, 6, 7, 8, 11, 12	3, 4, 7, 8, 11, 12	2, 4, 6, 8, 12	1, 2, 3, 4, 6, 7, 8, 11, 12	1, 2, 3, 4, 6, 8, 12	1, 2, 3, 4
6. Rh′Rh″ CdE/c	4, 7, 8, 9, 11, 12	4, 7, 8, 9, 11, 12	9, 11, 12	9, 11, 12	4, 7, 8, 11, 12	4, 7, 8, 11, 12	—	2, 4, 7, 8, 11, 12	2, 4	2, 4
7. Rh$_1$rh CDe/c	2, 4, 6, 8, 9, 11, 12	9, 11, 12	2, 4, 6, 8, 9, 11, 12	9, 11, 12	2, 4, 6, 8, 12	—	2, 4, 6, 8, 12	1, 2, 3, 4, 6, 8, 12	1, 2, 3, 4, 6, 8, 12	1, 2, 3, 4
8. Rh$_1$Rh$_2$ CDE/c	9, 11, 12	9, 11, 12	9, 11, 12	9, 11, 12	—	—	—	1, 2, 3, 4	1, 2, 3, 4	1, 2, 3, 4
9. Rh′Rh′ Cde/C	6, 7, 8, 9, 11, 12	7, 8, 9, 11, 12	6, 8, 9, 11, 12	9, 11, 12	6, 7, 8, 11, 12	7, 8, 11, 12	6, 8, 12	5, 6, 7, 8, 11, 12	5, 6, 7, 8, 12	5, 6, 7, 8
11. Rh$_1$Rh$_1$ CDe/C	6, 8, 9, 11, 12	9, 11, 12	6, 8, 9, 11, 12	9, 11, 12	6, 8, 12	—	6, 8, 12	5, 6, 7, 8, 12	5, 6, 7, 8, 12	5, 6, 7, 8
12. Rh$_1$Rh$_2$ CDE/C	9, 11, 12	9, 11, 12	9, 11, 12	9, 11, 12	—	—	—	5, 6, 7, 8	5, 6, 7, 8	5, 6, 7, 8

TABLE 4.—PROBABILITIES OF EXCLUSION OF PATERNITY FOR MEN OF VARIOUS PHENOTYPES

Phenotype	Frequency in Population	Probability of Exclusion
1. cde	0.1436	0.4495
2. cdE	0.0131	0.3628
3. cDe	0.0241	0.4414
4. cDE	0.1266	0.3355
5. Cde/c	0.0062	0.1845
6. CdE/c	0.0003	0.0973
7. CDe/c	0.3577	0.1066
8. CDE/c	0.1299	0
9. Cde/C	0.0001	0.5162
11. CDe/C	0.1973	0.4448
12. CDE/C	0.0011	0.4175
Unknown	1.0000	0.2500

Although this could be done in the present case, it results in algebraic expressions which are complicated and extremely long, and which are much less convenient to use than the numerical tables which can be prepared, from the formulas of table 1, for any population. Consequently, the more convenient procedure is to add from table 2, for a man of given phenotype, using table 3 as a guide, all children's frequencies which, for each type of mother, exclude paternity. This results in table 4.

The general chances of excluding the paternity of a falsely accused man in an English (or American) population are thus found to be 0.2500. The value arrived at by Fisher on the basis of approximate gene frequencies which were rounded to two decimals, namely 0.2520, is thus seen to be amazingly close. The individual probabilities for men of the various Rh phenotypes agree less well. For instance, for a man of type Cde/C Fisher gives 0.5801, as opposed to 0.5162 found above. The present method could of course be applied to any population.

Wiener (1950) studied 88 cases of exclusion of paternity, and found that the number of cases excluded by the Rh groups alone and by Rh and ABO and/or MN agreed well with the assumption of 25% exclusions by the Rh groups. Allen, Jones and Diamond (1954) have calculated (without, however, deriving general formulas) that if all seven Rh sera (anti-C, anti-C^w, anti-c, anti-D, anti-E, anti-e and anti-f) could be used the chances of exclusion would rise to 35%.

SUMMARY

A general derivation of the frequency with which children of various Rh phenotypes are born to mothers of the various phenotypes is given, and from this the probability is calculated of excluding the paternity of a falsely accused man in a population having Rh gene frequencies like those of the English. The method is applicable to any population.

REFERENCES

ALLEN, F. H., A. R. JONES AND L. K. DIAMOND 1954. Medicolegal applications of blood grouping. *N. E. J. Med.* 251: 146.

BOYD, W. C. 1954. Maximum likelihood method for estimation of gene frequencies from MNS data. *Am. J. Hum. Gen.* 6: 1.

BOYD, W. C. 1954b. Shortened maximum likelihood estimation of Rh gene frequencies. *Am. J. Hum. Gen.* 6: 303.

BOYD, W. C. 1955. The chances of excluding paternity by the MNS blood group system. *Am. J. Hum. Gen.* **7**, 199.

COTTERMAN, C. W. 1951. A note on the detection of interchanged children. *Am. J. Hum. Gen.* 3: 363.

FISHER, R. A. 1946. The fitting of gene frequencies to data on *rhesus* reactions. *Ann. Eugenics* 13: 223.

FISHER, R. A. 1947. Note on the calculation of the frequencies of *rhesus* allelomorphs. *Ann. Eugenics* 13: 224.

FISHER, R. A. AND R. R. RACE 1946. Rh gene frequencies in Britain. *Nature* 157: 48.

RACE, R. R. AND R. SANGER 1950, 1954 *Blood Groups in Man.* Blackwell Sci. Pub. Oxford.

WIENER, A. S. 1950. Heredity of the Rh blood types IX Observations in a series of 526 cases of disputed parentage. *Am. J. Hum. Gen.* 2: 177.

WIENER, A. S. 1952. Heredity of the M-N-S blood types. Theoreticostatistical considerations. *Am. J. Hum. Gen.* 4: 37.

WIENER, A. S., M. LEDERER AND S. H. POLAYES 1930. Studies in Isohemagglutination: IV On the chances of proving non-paternity: with special reference to blood groups. *J. Immunol.* 19: 259.

ZARNIK, B. 1930. Vjerojatnost pozitivnog nalaza kod ispivanja krvnih grupa u svrhu iskljucivanja ocinstva. *Medicinski Pregled 5:* 1.

A Genetic Study of Multiple Polyposis of the Colon (With an Appendix Deriving a Method of Estimating Relative Fitness)[1]

T. EDWARD REED AND JAMES V. NEEL

Heredity Clinic, Institute of Human Biology, University of Michigan, Ann Arbor, Michigan

MULTIPLE POLYPOSIS of the colon is a condition characterized by the occurrence of numerous polyps throughout the colon and/or rectum. The polyps usually make their appearance during the first and second decades of life, but occasionally may not arise until the fourth decade or possibly even later. Although the polyps themselves may be associated with symptoms referable to the large bowel, the disease is of clinical significance primarily because of the tendency of one or more of the polyps to become malignant, with death from carcinoma of the colon or rectum at a relatively early age. The disease has long been known to have a familial distribution (cf. Cripps, 1882), with Cockayne (1927) apparently the first to point out that this distribution was characteristic of a trait dependent on a dominant gene. Dukes (1952) has provided an excellent review and bibliography of the disease.

Approximately 10 per cent of all adults can be demonstrated by combined sigmoidoscopic and X-ray studies to possess one or more polyps of the colon (cf. Helwig, 1947; Swinton and Haug, 1947; Bacon, 1949; Gianturco and Miller, 1953). However, only a small fraction of these persons has true multiple polyposis. When appropriate diagnostic studies are carried out, there is seldom any problem involved in differentiating between the person who has hundreds or even thousands of polyps—and has multiple polyposis—and the person who has two or three, or even five polyps, but does not have the multiple polyposis with which this study is concerned.

The two diseases with which multiple polyposis can most readily be confused are the so-called Peutz-Jeghers syndrome of diffuse intestinal polyposis and abnormal pigmentation (cf. Jeghers, McKusick, and Katz, 1949), and the syndrome of polyposis of the colon associated with osteomatosis and fibromatosis (Gardner and Richards, 1953). Both of these diseases are apparently much rarer than classical multiple polyposis, and were not encountered in the course of this study.

The present investigation was undertaken in an effort to develop a more rounded picture of the genetics of this disease than is currently available. More specifically, in addition to accumulating further data concerning the inheritance of this condition, we have attempted to evaluate the biological handicap it imposes on affected persons,

Received May 20, 1955.

[1] This study was supported in part by a grant from the U. S. Atomic Energy Commission. We are indebted to Drs. H. M. Pollard, R. J. Bolt, and B. I. Hirschowitz of the Department of Internal Medicine of the School of Medicine for their generous cooperation in certain aspects of this investigation. We are also indebted to Dr. Norman Ewing, Vincennes, Indiana, for a large amount of information concerning Kindred 1554, and to Dr. R. Sidell, Grand Rapids, Michigan, for calling our attention to the propositus in Kindred 2067. Finally, we have profited greatly from discussion with our colleagues Dr. R. S. Krooth and Dr. W. J. Schull.

and to estimate the frequency of the trait in the general population. Assuming genetic equilibrium, these data permit certain preliminary speculations concerning the rate with which the gene or genes responsible for this trait are appearing through mutation. Although, as will be pointed out in due course, large errors are inevitable in certain of these calculations, it is felt that to the extent that they contribute to arriving at an order of magnitude for a basic biological phenomenon, the calculations are worthwhile and of general interest.

GENETIC STUDIES ON MULTIPLE POLYPOSIS

The 23 kindreds on which genetic studies have been carried out were located in the following ways:

1. A survey of the records of the University Hospital of the University of Michigan for the period 1935–1950, which yielded 16 families for study. Subsequent experience has revealed that by no means all cases of polyposis seen at the University Hospital during this period were coded as such and that, further, some cases properly coded as polyposis were overlooked in the initial survey. However, so far as is known, no element of bias entered into the selection of these particular kindreds.

2. Correspondence with a number of Michigan physicians specializing in gastroenterology, which yielded three kindreds.

3. Systematic follow-up of all death certificates filed with the state of Michigan during 1950–52 inclusively on which the cause of death is listed as carcinoma of the colon or rectum and the individual was below the age of 40. This procedure, undertaken in an effort to estimate the frequency of the trait (see below), yielded four kindreds.

BASIC DATA

The basic data on the 23 kindreds studied are presented in Tables 1 and 2. Each kindred contains at least one medically diagnosed case of multiple polyposis. One hundred and nine affected or possibly affected individuals are described. Seventy of these are definitely known to have had multiple polyposis. Thirteen are known to have developed cancer of the colon or rectum and, because of their close biological relationship to an individual known to have polyposis, are assumed also to have had polyposis. The 26 remaining persons are included because of lay reports or inconclusive medical reports of polyposis, bowel cancer, or significant bowel complaints, such as rectal bleeding. There are 65 males and 44 females, and 75 of the 109 individuals were residents of the state of Michigan at the time of investigation or at death.

The kindreds can be conveniently classified into two groups according to the existence or absence of good evidence for the presence of two or more affected individuals in a kindred. This division yields 14 kindreds of the familial type and 9 which are not clearly familial. Pedigrees and data concerned with the 14 familial kindreds are presented in Figure 1 and Table 1; data on the remaining 9 kindreds are given in Table 2. As Table 2 shows, several, or perhaps most, of these 9 kindreds may well contain two or more affected persons but data required for decision are lacking. Poor cooperation from close relatives of the propositus accounts for much of this

TABLE 1.—DATA ON AFFECTED AND POSSIBLY AFFECTED MEMBERS OF 23 KINDREDS, EACH KINDRED CONTAINING AT LEAST ONE MEDICALLY DIAGNOSED CASE OF MULTIPLE POLYPOSIS OF THE COLON

Key: + = Yes; − = No or none; ? = Diagnosis made but age unknown; • = No information; P = Propositus; "definite" = Data from hospital record, patient's physician, or death certificate.

Kindred	Individual	Sex	Michigan resident	Type of Evidence				Age at				Comments
				Definite polyposis with or without cancer	Definite primary large bowel cancer; polyposis not known	Reliable lay report of polyposis or large bowel cancer	Dubious lay report of significant intestinal symptoms, polyposis or large bowel cancer	First symptoms	Diagnosis of polyposis	Diagnosis of primary large bowel cancer	Death	
A. Kindreds certainly or very probably containing two or more cases of multiple polyposis												
809	I-1	M	−	−	−	−	+	•	•	•	44	Son reports I-1 underwent abdominal surgery for unknown cause 6 mos. prior to death; hospital record unavailable.
	II-1	F	+	+	−	−	−	32	34	34	34	
	II-2	F	−	+	−	−	−	45	47	47	49	
	II-5	M	+	+	−	−	−	30	42	42	−	
	III-1(P)	M	+	+	−	−	−	25	25	32	−	Underwent subtotal colectomy with ileosigmoidostomy at age 26. Rectal polyp removed age 32 thought to be malignant. Asymptomatic when last seen age 33.
832	I-1	M	−	−	+	−	−	•	−	67	70	At operation for carcinoma of sigmoid colon at age 67, history obtained of other bowel surgery 17 yrs. previously—?carcinoma. Operation age 67 only palliative colostomy. Polyposis not diagnosed.
	II-5	M	−	−	−	−	+	•	•	•	•	Reported by mother, on dubious grounds, to have had polyposis.
	II-7(P)	M	+	+	−	−	−	43	44	44	45	
	II-8	F	−	+	−	−	−	38	41	41	42	
	III-22	M	+	+	−	−	−	20	20	−	−	No symptoms; studied because of disease in father.
	III-24	M	+	+	−	−	−	−	14	−	−	No symptoms; studied because of disease in father.
	III-25	M	+	+	−	−	−	−	11	−	−	No symptoms; studied because of disease in father.
	III-28	F	−	+	−	−	−	−	23	−	−	

1554	I-2	F	−	−	−	+	−	•	•	•	50	Son reports I-2 died of large bowel cancer at age 50. The physician of II-17, 18, and 22 confirms this report. Hospital record unavailable.
	I-4	M	−	−	−	+	−	•	•	•	42±	Nephew and physician of II-17, 18, and 22, familiar with family, report I-4 died of large bowel cancer. Detailed medical record not available.
	I-7	M	−	−	+	−	−	60	−	60	62	Found to have inoperable carcinoma of colon; underwent only palliative colostomy. Polyps not reported on sigmoidoscopy or barium enema.
	II-6	M	−	+	−	−	−	•	36	37	−	
	II-8	F	−	+	−	−	−	29	31	29	40	
	II-9	M	−	−	−	+	−	•	•	•	33±	Brother and physician of II-17, 18, and 22 report cancer of colon. Detailed medical records unavailable.
	II-17	M	−	+	−	−	−	40	40	−	−	
	II-18	F	−	+	−	−	−	•	37	−	−	
	II-19	M	−	−	−	−	+	•	−	−	−	According to physician of II-17, 18, and 22, II-19 had rectal bleeding at about age 41. No examination is known to have been made.
	II-20	M	−	+	−	−	−	33	33	−	−	
	II-21	F	−	−	+	−	−	29	−	29	29	Pathology report on resected sigmoid colon fails to mention polyposis.
	II-22	F	−	+	−	−	−	•	32	33	33	
	II-23	M	−	+	−	−	−	−	27	28	29	
	II-25(P)	M	+	+	−	−	−	24	24	−	−	
	III-12	F	−	+	−	−	−	−	27±	−	−	
1801	I-1	M	+	−	+	−	−	•	−	?	37	Father of I-1 said to have died at about age 35, cause unknown to informants. Mother of I-1 died about age 60. No information about sibs of I-1.
	I-3	M	+	−	−	−	−	−	−	−	81	I-3 had no gastrointestinal complaints but 1 of his 6 sibs is said to have died at age 66±, possibly of "cancer of the liver." Father of I-3 is known to have had bloody diarrhea and a rectal mass before his death at age 74.
	I-4	F	+	−	−	−	−	−	−	−	−	I-4 was in good health at age 72 but had a half-sister who is known to have died at age 59 of cancer of the large bowel.
	II-2	M	+	−	−	−	−	•	−	−	37	Surgery revealed neoplasm involving stomach and transverse colon, thought by surgeon to be primary in stomach.
	II-3	F	+	+	−	−	−	•	55	55	−	Sibs II-4, 6, and 7 were medically examined and found normal at ages 58, 57, and 53 respectively. II-5 died at 6 mos.

TABLE 1.—*Continued*

Kindred	Individual	Sex	Michigan resident	Type of Evidence				Age at				Comments
				Definite polyposis with or without cancer	Definite primary large bowel cancer; polyposis not known	Reliable lay report of polyposis or large bowel cancer	Dubious lay report of significant intestinal symptoms, polyposis or large bowel cancer	First symptoms	Diagnosis of polyposis	Diagnosis of primary large bowel cancer	Death	
A. Kindreds certainly or very probably containing two or more cases of multiple polyposis—*Continued*												
1801	III-9(P)	F	+	+	−	−	−	15	15	17	18	
	III-10	M	+	+	−	−	−	13	13	14	19	
	III-11	F	+	−	+	−	−	8	−	9	9	
	III-12	F	+	+	−	−	−	−	28	−	−	Diagnosed during this study
1824	I-1	M	+	−	−	+	−	•	•	•	45	Hospital record of II-1 states that I-1 died of cancer of the rectum.
	II-1	M	+	−	+	−	−	26	−	27	27	Polyp removed on biopsy; when cancer of rectum diagnosed, only palliative colostomy performed.
	III-1(P)	F	+	+	−	−	−	−	23	−	−	
1826	I-1	M	+	−	−	−	+	•	−	−	54±	Reported by a daughter-in-law and a granddaughter-in-law to have died of cancer of bowel in 1896.
	II-1	F	−	−	−	+	−	•	−	−	40±	Reported by a daughter and a sister-in-law to have died of cancer of bowel.
	II-5	M	+	−	+	−	−	•	−	46	47	Death certificate filed in 1916 lists cancer of lower bowel, bladder, and kidney as cause of death.
	II-7	F	+	+	−	−	−	•	40	40	40	
	II-8	M	+	+	−	−	−	41	44	−	46	Presented with pelvic abscess, which was drained. Sigmoidoscopic and X-ray studies revealed obstructive lesion in colon; patient died without definitive surgery.
	III-1	F	−	−	−	+	−	•	−	−	30	Reported by a brother, sister, and a sister-in-law to have died following diarrhea and abdominal cramps. Death certificate dated 1911 gives cause of death as tuberculosis of the bowels.
	III-2	M	−	−	−	−	+	•	−	−	33	Reported by a brother, sister, and a sister-in-law to have died following diarrhea and cramps. Death certificate dated 1921, however, gives pulmonary tuberculosis as the cause of death.

	III-3	M	−	−	−	+	−	•	−	−	39	Reported by a brother, sister, and a sister-in-law to have died following diarrhea and cramps. Death certificate dated 1925 gives cause of death as tubercular peritonitis.
	III-4	F	−	−	−	−	+	•	−	−	30	Reported by a brother, sister, and sister-in-law to have died following intestinal complaints. No medical record available.
	III-5	F	+	+	−	−	−	57	58	−	−	First diagnosed during this study.
	III-7(P)	M	+	+	−	−	−	37	37	37	51	One of the 2 separately ascertained propositi in this kindred. Had subtotal colectomy at 37.
	III-21(P)	M	+	+	−	−	−	50	50	50	51	One of the 2 separately ascertained propositi in this kindred.
	III-23	M	+	+	−	−	−	•	51	?	−	Cancer of the cecum was diagnosed some time before diagnosis of polyposis was made.
	III-27	M	+	−	−	−	+	•	−	−	35	Hospital report states III-27 died shortly after arrival, apparently of general peritonitis. There was no autopsy.
	III-29	M	+	−	+	−	−	39	−	39	−	Presented with symptoms of large bowel obstruction. At operation a malignancy involving splenic flexure was resected. Pathology report fails to mention polyposis.
	III-30	F	+	+	−	−	−	26±	38	38	38	
	III-34	M	+	−	−	−	+	•	−	−	−	Examined twice following discovery of multiple polyposis in sibs. First examination negative; second exam (barium enema) revealed solitary polyp at hepatic flexure.
	III-35	F	+	+	−	−	−	•	31	32	33	
	III-36	M	+	+	−	−	−	•	32	32	−	Examined following death of sister.
	III-37	M	+	+	−	−	−	−	31	31	−	Examined following death of sister.
	III-40	F	−	+	−	−	−	−	38	−	−	Diagnosed during this study.
	III-41	M	+	+	−	−	−	−	35	−	−	Diagnosed during this study.
	IV-9	M	+	+	−	−	−	−	19	−	−	Diagnosed during this study.
	IV-24	M	+	+	−	−	−	−	22	−	−	Diagnosed during this study.
	IV-26	F	+	+	−	−	−	−	17	−	−	Diagnosed during this study.
	IV-27	M	+	+	−	−	−	−	15	−	−	Diagnosed during this study.
	IV-28	F	+	+	−	−	−	−	12	−	−	Diagnosed during this study.
1861	I-2	M	+	−	−	−	−	−	−	−	−	I-2 described himself as being healthy at age 47. His mother died at 30 from unknown causes. His sister is healthy at 45.
	I-3	F	+	−	−	−	−	−	−	−	−	I-3 described herself as being in fair health at age 44. A hospital report states that at age 43 X-rays of the colon were normal. Her 10 sibs are reported normal. Father is said to have died at 52 from heart trouble, mother at about 63 following an operation for hernia.

TABLE 1.—*Continued*

Kindred	Individual	Sex	Michigan resident	Type of Evidence				Age at				Comments
				Definite polyposis with or without cancer	Definite primary large bowel cancer; polyposis not known	Reliable lay report of polyposis or large bowel cancer	Dubious lay report of significant intestinal symptoms, polyposis or large bowel cancer	First symptoms	Diagnosis of polyposis	Diagnosis of primary large bowel cancer	Death	
A. Kindreds certainly or very probably containing two or more cases of multiple polyposis—*Continued*												
1861	II-1(P)	F	+	+	−	−	−	10	10	−	10	Presented with bloody diarrhea. Died of a transfusion reaction following subtotal colectomy.
	II-3	M	+	+	−	−	−	9	9	−	−	Was in good health at age 19. No surgery has been performed.
1862	I-1	M	+	−	−	+	−	•	•	•	60	Reported by 5 of his children to have died in 1919 of cancer of the intestinal tract following symptoms including massive rectal bleeding. Hospital record and death certificate not available. Had 9 sibs, 3 of whom are reported to have died from cancer of the colon. The parents of I-1 are both reported to have died in their 40's.
	II-1	M	−	−	+	−	−	•	•	?	40	Death certificate lists cancer of colon as cause of death; further details unobtainable.
	II-2	F	+	−	+	−	−	•	−	?	49	Death certificate lists cancer of rectum as cause of death; further details unobtainable.
	II-4	F	+	−	−	−	+	•	−	−	37	Several sibs report that II-4 had much rectal bleeding but death certificate gives cause of death as carcinoma of the spleen.
	II-12(P)	F	+	+	−	−	−	25	25	32	34	
	III-26	M	−	+	−	−	−	−	21	−	−	Polyposis detected during examination because of anal condylomas.
1928	I-1	M	+	−	+	−	−	•	−	?	59	Reported by his daughter to have undergone abdominal surgery, resulting in a "short circuit" of the intestine. Hospital records not available. Death certificate gives cause of death as cancer of the rectum.
	II-1	F	+	+	−	−	−	30±	38	−	39	Following several months of bloody diarrhea and abdominal cramps, developed a pelvic abscess, which was immediate cause of death. X-rays showed persistent narrowing of sigmoid colon. Autopsy report describes abscess and polyposis but does not mention neoplasm.

	II-2	M	+	+	−	−	−	•	42	−	42	Died following subtotal colectomy because of intractable diarrhea. No malignancy of colon.
	II-5	F	+	+	−	−	−	26	44	−	44	Died following 18 yrs. of intermittently bloody diarrhea. Autopsy revealed multiple polyposis without malignant degeneration.
	III-1(P)	F	+	+	−	−	−	17	25	−	−	Polyposis diagnosed during work-up for chronic diarrhea.
1982	I-1	M	−	−	+	−	−	•	−	44	47	The wife of I-1 states that his mother died of cancer and the wife of II-1 says that I-1's mother died of cancer of the intestine. This second informant also states that I-1 had several sibs who died of "intestinal ailment." I-1 underwent resection at age 44 for carcinoma of descending colon; pathology report fails to mention polyposis. Second (?) malignancy detected 3 yrs. later, with death shortly thereafter.
	II-1(P)	M	+	+	−	−	−	30	30	30	32	
	II-2	M	−	+	−	−	−	•	36	−	−	
	III-1	F	+	+	−	−	−	•	19	−	−	Asymptomatic; diagnosed during this study.
2070	I-1	M	−	−	−	+	−	•	•	•	39	When admitted to hospital, II-2 stated that I-1 died from cancer of the bowel, that the mother of I-1 died from cancer involving the pelvic organs, and that at least one sib died from cancer of the bowel. No hospital record on I-1 is available.
	II-2	M	+	+	−	−	−	32	32	32	42	Suicide was cause of death.
	III-1(P)	M	+	+	−	−	−	17	17	−	−	Diagnosis was made following rectal bleeding.
3361	I-2	F	+	−	−	−	−	•	−	−	73	Hospitalized shortly before death with findings of partial intestinal obstruction. No operation was performed.
	II-2	F	+	−	−	−	−	•	−	−	35	Exploratory laparotomy shortly before death revealed multiple metastatic nodules in liver. Pathologist's report suggests primary in gastro-intestinal tract.
	II-4	F	+	−	+	−	−	41	−	41	−	On two occasions a rectal polyp was removed; last specimen reported as showing malignant degeneration. However, diagnosis of multiple polyposis not made.
	II-5(P)	F	+	+	−	−	−	24	39	39	−	
3496	I-2	F	+	+	−	−	−	−	49	−	−	An orphan with no knowledge of parents or sibs; polyposis diagnosed during this study.
	II-2(P)	M	+	+	−	−	−	28	30	30	30	
	II-5	F	+	+	−	−	−	−	23	−	−	Diagnosed during this study.

TABLE 1.—*Concluded*

Kindred	Individual	Sex	Michigan resident	Type of Evidence: Definite polyposis with or without cancer	Type of Evidence: Definite primary large bowel cancer; polyposis not known	Type of Evidence: Reliable lay report of polyposis or large bowel cancer	Type of Evidence: Dubious lay report of significant intestinal symptoms, polyposis or large bowel cancer	Age at: First symptoms	Age at: Diagnosis of polyposis	Age at: Diagnosis of primary large bowel cancer	Age at: Death	Comments
A. Kindreds certainly or very probably containing two or more cases of multiple polyposis—*Continued*												
3498	I-1	M	−	+	−	−	−	•	46	41	46	According to I-1, his mother and 3 of his sisters died from cancer of the rectum. This is confirmed by I-1's daughter, II-5, who also states that a brother of I-1 has had a colostomy.
	II-1	M	−	−	−	+	−	•	•	•	−	Son of an apparently normal brother of I-1. According to II-5, II-1 had 6 rectal polyps removed at age 29±.
	II-3(P)	M	+	+	−	−	−	30	30	33	33	
	II-5	F	−	+	−	−	−	•	26±	−	−	
B. Kindreds not certainly or very probably containing two or more cases of multiple polyposis												
1825		F	+	+	−	−	−	5	8	−	−	
1963		M	+	+	−	−	−	•	27	28	−	
2067		M	+	+	−	−	−	5	5	−	−	See text.
2068		M	+	+	−	−	−	25	28	−	−	
2069		M	+	+	−	−	−	•	27	−	−	
2071		M	+	+	−	−	−	•	23	23	31	
4029		F	+	+	−	−	−	23	25	25	26	
4057		M	+	+	−	−	−	28	30	−	−	
4103		M	+	+	−	−	−	39	39	39	39	

uncertainty. Particular aspects of Tables 1 and 2 are considered in subsequent sections.

The familial kindreds are represented in the pedigrees and table as far as the data seem to warrant. It should be emphasized that the nature of the trait is such that pedigrees alone represent only a portion of the pertinent data; for more complete information the table should be consulted. Two of the 14 familial kindreds are of particular interest. Kindred 1826 has recently been described by Neel, Bolt, and Pollard (1954) and is noteworthy for its size, including 17 medically diagnosed cases of multiple polyposis. Another kindred, 1801, is remarkable for having a sibship of four persons (III, 9–12), all of whom have or have had diagnoses of multiple polyposis and/or cancer of the large bowel, three of them dying of cancer of the bowel under the age of 20 (at ages 9, 18, and 19). The earliest age at death from cancer of the large bowel among the other 22 kindreds is 26. The other unusual feature of this kindred is that in addition to diagnosed multiple polyposis in the mother of this sibship, the family physician reports that the father, II-2, was found at exploratory laparotomy to have cancer involving the stomach and transverse colon, appearing to be primary in the stomach, and the father's father had cancer of the colon. Both men died at age 37. In addition, the father's paternal grandfather is reported to have died at about the age of 35 of unknown causes. Since in none of the other 13 familial kindreds is there an affected sibship having one parent with multiple polyposis and the other with a history comparable to the above mentioned one, the question arises whether this parental history is related to the three early deaths in the sibship. It is conceivable that the father's cancer was primary in the colon like that of his own father, both arising from multiple polyposis, in which case some or all of the three early deaths in sibship III, 9–12 may have been of persons homozygous for the polyposis gene. This conjecture can neither be proved nor disproved at present.

One kindred (2067) of the nine described in Table 2 also deserves special comment because it illustrates the diagnostic difficulties which occasionally arise. At age 6 the propositus came to medical attention because of a mass protruding from the rectum; this was found to be a prolapsed polyp. After sigmoidoscopic and X-ray studies, a surgeon made a diagnosis of multiple polyposis and performed a hemicolectomy with anastomosis of the mid-transverse colon and distal sigmoid. The pathologist's report on the specimen reads as follows: "Specimen consists of a 40 cm. segment of colon with attached mesentery and a separate pedunculated polypoid granular reddish-grey lesion about 1 cm. in gross diameter. On section, the wall is essentially normal in thickness. The mucosa of the specimen contains 5 reddish-grey granular lesions varying from 5 to 10 mm. in gross diameter. Three of the specimens have very long, soft pedicles. . . ." Microscopic sections of these lesions were typical of polyps of the colon. The boy was seen by us at age 10; sigmoidoscopic examination revealed no polyps, and two barium enemas, although demonstrating a shortened colon, likewise failed to provide evidence for polyposis. A brother, aged 7, his father, aged 38, and his mother, aged 31, were all negative to sigmoidoscopy and barium enema. The maternal grandmother underwent an exploratory laparotomy at about age 60 and was found to have "generalized metastatic adenocarcinoma of the abdominal cavity" (hospital report); a maternal great aunt is reported by a physician to have undergone surgery because of carcinoma of the bowel at age 67. While there is no doubt that the propositus had multiple polyps of his distal colon, the complete absence of polyps in the remaining large bowel at the time of our examination raises doubt as to whether this is the type of multiple polyposis with which this study is otherwise concerned. It should be noted that the inclusion of this dubious case in the series does not affect any of the calculations to be presented below.

TABLE 2. FURTHER DATA ON KINDREDS IN WHICH NO MEMBERS IN ADDITION TO THE PROPOSITUS ARE KNOWN OR STRONGLY BELIEVED TO HAVE MULTIPLE POLYPOSIS

(Propositi are described in Table 1-B.)

Key: L = Living at time of investigation; D = Deceased at time of investigation, from non-violent cause; K = Deceased at time of investigation, from violent, accidental cause.

Kindred	Parents of propositi (Age at time of investigation or at death)		Sibs of propositi		Significant intestinal symptoms in near relatives of the propositus	Details (All relationships refer to propositus)
	Father	Mother	No. living to age 5 or more (full & half-sibs)	Age range		
1825	K36	L35	1	10	+	Father had good health until accidental death. Mother in good health; normal sigmoidoscopic examination at 35. Father had 2 sibs, now aged 43 and 36, the older found to have cancer of recto-sigmoid colon at age 41, which was resected. Barium enema, sigmoidoscopy, and pathology reports on this man did not mention multiple polyposis. Paternal grandfather was drowned at about 30; paternal grandmother, according to death certificate, died at 55 from cancer of cervix. Mother has one sib, normal at 29. Maternal grandfather in good health at 59; maternal grandmother at age 55 reports having had "colon trouble" for 15 yrs. but sigmoidoscopy revealed no polyps.
1963	L69	L69	9	25–40	−	7 full sibs, 2 half-sibs. Sibs and parents live out of Michigan. Mother writes that she, father, and all sibs are in good health.
2067	L38	L31	1	7	+	See text.
2068	L58	L51	3	25–32	+	(All information from the propositus): Sibs appear to be normal. Father in good health no bowel symptoms known among his sibs and parents. Mother has had "stomach and bowel upsets" for some time. Sibs and parents refuse examination. Mother has 7 sibs, one of whom is reported to have had a colectomy like that of propositus; he refuses to allow his medical record to be examined. No other persons with gastro-intestinal complaints are known.
2069	L55	L51	2	26–32	+	(All information from propositus and one sib): Sibs and parents appear normal. Father has 4 sibs, one of whom is reported to have died at 63± from intestinal obstruction. This report cannot be confirmed or refuted. No other persons with g.i. complaints are known.

2071	L56	L54	5	21–33	–	One sib killed in accident at 21. Two sibs examined at ages 30 and 29 were normal by sigmoidoscopy and barium enemas. Fairly detailed family history through grandparental generation reveals no person with significant intestinal symptoms.
4029	L60	L57	5	16–35	+	One half-sib aged 35 normal to sigmoidoscopy and barium enema. Three full sibs, ages 23, 21, and 18 at time of examination; barium enema on the oldest, barium enema and sigmoidoscopy on the middle sib, and sigmoidoscopy on the youngest all reveal normal colons. The youngest full sib, not examined, appears normal at 16. The parents do not have significant g.i. complaints but each is vaguely reported to have a sib with possible intestinal symptoms. No other persons with significant intestinal complaints are known.
4057	K53	L50	2	25–32	+	Father normal; killed in accident. Mother normal. Older sib normal by sigmoidoscopy and barium enema. Younger sib has had some rectal bleeding but will not be examined. No other persons with g.i. complaints are known.
4103	D48	L64	1	33	–	Father probably died of heart disease. Mother well. Sib well. No relatives with g.i. complaints known.

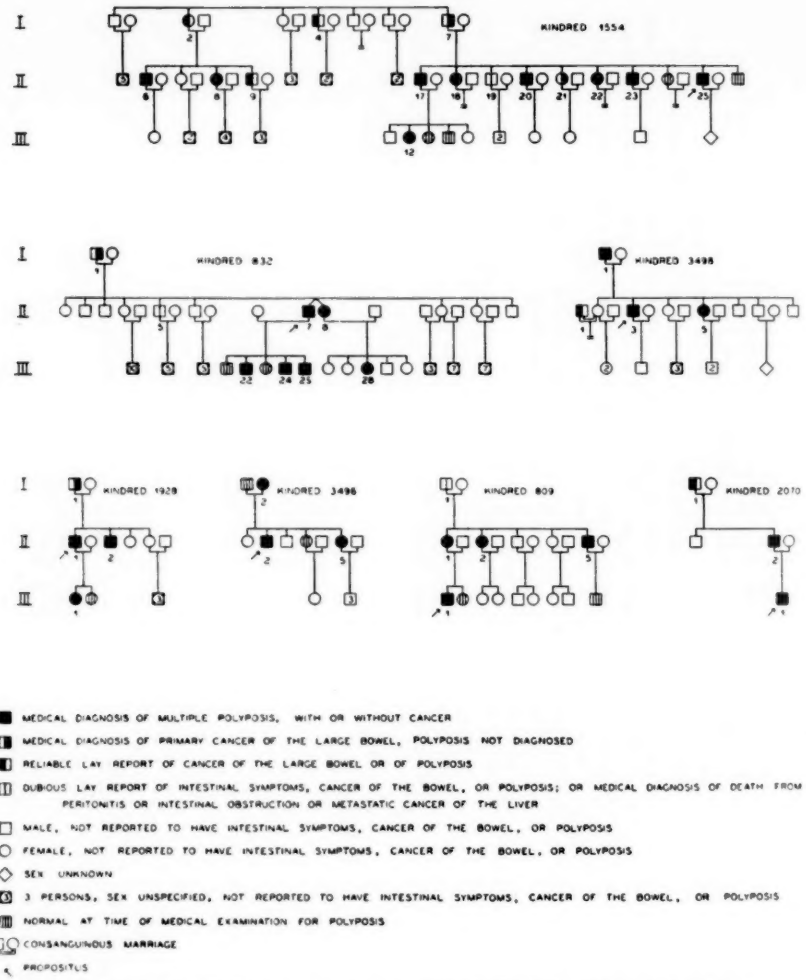

FIG. 1. Pedigrees of kindreds which contain two or more individuals with multiple polyposis.

INHERITANCE

The published pedigrees of multiple polyposis clearly indicate that this trait is usually, if not always, determined by a dominant gene of fairly high penetrance. The distribution of affected persons within the kindreds of the present study is in keeping with the reports in the literature, medically diagnosed polyposis occurring in two generations of 10 kindreds and in three generations of one.

Unfortunately, a precise calculation of the proportion of affected and unaffected sibs within sibships having a parent with polyposis, desirable as a check on the

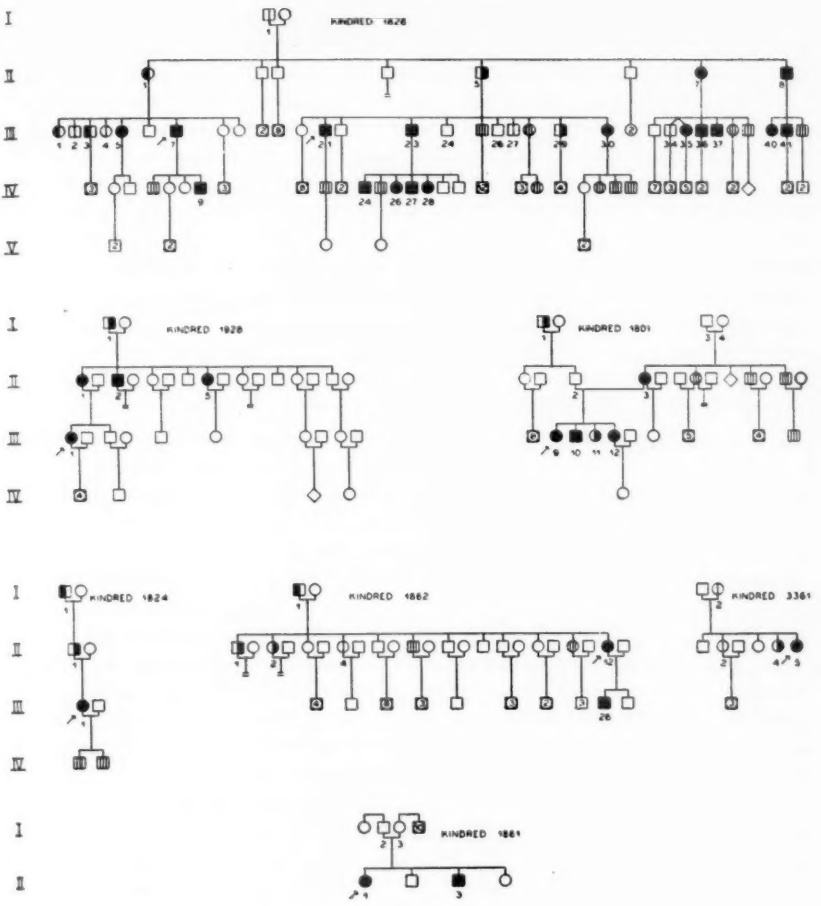

FIG. 1

hypothesis of dominant inheritance and on the degree of penetrance under this hypothesis, is not possible. The late onset of symptoms and diagnosis in some individuals (about 10 per cent of the medically diagnosed cases are first diagnosed over age 45) and the lack of diagnostic information on a number of individuals account for the fact that not one sibship in the 14 familial kindreds gives critical information on the segregation of the gene for polyposis. However, it is possible to obtain a rough test for agreement with the expected 1:1 ratio in segregating sibsips by using only sibships in which persons reported to be normal are well within the age of manifestation of the trait, and, at the same time, employing conservative criteria in classifying individuals with respect to the trait. In this calculation, two kinds of sibships—that of the propositus and, where possible, that of his affected parent—have been utilized;

the propositus and affected parent have been excluded from the calculation. Such a test, in which the minimum age of "normal" sibs was set at 41 (except for one individual, Kindred 1826, III-24 described below), and in which sibs reported as affected by "reliable" lay sources were counted affected, and sibs dubiously reported affected were counted normal, was applied to 9 sibships (Kindred 809, II, 1–5; Kindred 832, II, 1–11; Kindred 1554, 1, 1–7; Kindred 1554, II, 6–9; Kindred 1801, III, 9–12; Kindred 1826, II, 1–8; Kindred 1826, III, 1–9; Kindred 1826, III, 20–30, counting III, 24 who died at 32 of "epilepsy" as normal; Kindred 2070, II, 1–2). It showed 24 affected persons and 36 unaffected, a proportion not differing significantly from 1:1. If there is 1:1 gene segregation and reduced penetrance, the observed proportion deviates from 1:1 in the direction expected. It seems a reasonable assumption that multiple polyposis is determined by a single gene whose penetrance is of the order of 90 per cent in persons of age 50, higher in still older persons.

FREQUENCY OF MULTIPLE POLYPOSIS

An attempt was made to obtain an estimate of the frequency at birth of individuals heterozygous for the gene for multiple polyposis. This estimate cannot be directly determined, but an approximate estimate can be based on the proportion of individuals dying within some time-interval who have had multiple polyposis. This approximation is biased toward underestimating the frequency at birth because some persons with the gene for polyposis will not be recognized at their death, either dying of cancer secondary to polyposis without diagnosis of polyposis, or from some other cause. A further bias, but in the opposite direction, exists when applying this method to American populations. This second bias is a consequence of the increasing absolute number of births per year in the U.S.A. and the decreased life expectancy of individuals with the gene. Its effect is to over-estimate the true frequency at birth. It is not possible to make any adequate allowance for these biases, but the former would seem to be more important. This estimate is probably an underestimate.

The frequency estimate was obtained in the following manner. The Department of Health of the state of Michigan furnished copies of the death certificates of all persons who died in Michigan before age 40 during the three-year period 1950–52 from primary carcinoma of the colon or rectum. One hundred and two such certificates were on file. On 25 certificates it was stated that an autopsy had been performed. The findings on these certificates were accepted as final. An effort was made to contact the next-of-kin of each of the remaining 77 deceased persons to obtain permission for the release of medical information to the Heredity Clinic. After receiving such permission, letters were written to the deceased's physician and to the hospitals where the deceased was studied, requesting copies of his medical records in order to determine whether he had multiple polyposis. In 59 cases medical reports were obtained, and in 18 they were not obtained. This procedure yielded 4 persons with multiple polyposis (Kindreds 3496, 3498, 4029, and 4103); only one (Kindred 4103) of these 4 persons had a death certificate which failed to state that multiple polyposis was present. The proportion of known cases of multiple polyposis is therefore 4/102 or 0.039 ± 0.019.

It soon became apparent that a frequency estimate based on this proportion would

be a minimal one. Thus, some death certificates fail to mention multiple polyposis even when the persons' hospital records do. Furthermore, in 3 of the 23 kindreds included in this study, the hospital records of persons who were parents of two or more children who themselves had polyposis, stated only that cancer of the large bowel was present. It is almost certain that these three persons (see Pedigrees 832, 1564, and 1982) had the gene and trait of multiple polyposis. In order to obtain a more reliable estimate of the proportion of persons dying before age 40 from primary cancer of the large bowel secondary to multiple polyposis, a survey was made of all the records of persons dying before age 40 from primary cancer of the rectum and colon who were studied at the University Hospital, Ann Arbor, in the period 1935–1944. Of 58 such persons, 6 had definite multiple polyposis and 1 had questionable multiple polyposis. Counting only the 6 definite cases, the proportion is 6/58 or 0.103 ± 0.040. It is clear that this may be an underestimate of the true proportion.[2]

In a population in equilibrium, the frequency (f) at birth of individuals with the gene for polyposis is equal to the frequency, among all deaths in a specified time interval, of individuals dying with the gene. If, for a specified population,

T = the specified time interval
P = the number of individuals with the gene for polyposis dying in T
D = the total number of deaths of all individuals in T
a = the observed proportion of individuals, among those dying before age 40 from primary cancer of the large bowel, whose cancer is secondary to multiple polyposis
b = the observed number of individuals dying before age 40 from primary cancer of the large bowel in T
c = the observed proportion of individuals, among those dying from cancer of the large bowel secondary to multiple polyposis, who die before age 40

then an approximation to P is given by ab/c. This estimate of P will be too low since some persons with the gene for polyposis fail to die from, or are not recognized as having died from, cancer secondary to multiple polyposis. For T we have used the three-year interval 1950–52. From the preceding section a is 0.103 ± 0.040; b for the state of Michigan is 102; c from Table 6 is 45/91 or 0.495 ± 0.052; D for the state of Michigan is 175,842. The frequency at birth is then approximated as

$$f = \frac{P}{D} \sim \frac{ab}{cD} = \frac{(.103 \pm .040)(102)}{(.495 \pm .052)(175{,}342)}$$

$\sim (1.21 \pm 0.49) \times 10^{-4}$ or 1 in 8,300 individuals. The expression ab/c neglects individuals with the gene for polyposis who die of causes other than cancer secondary

[2] An observation which may be pertinent here is that the frequency distribution of age at death from cancer of the large and small intestine (almost entirely due to the large intestine) shows a noticeable "bump" at the 30–34 year interval relative to the corresponding distribution for cancer of the stomach (data from Michigan Department of Health over the interval 1933–1945 inclusive.) Multiple polyposis will contribute to the former distribution but not to the latter so that it seems possible that the early age at death subsequent to polyposis may be responsible for this bump. If such is the case and if, in this age range, and in the absence of polyposis, the two frequency curves are proportional, one can estimate that cancer following polyposis may account for as much as a quarter of all cancer of the colon under age 40.

to polyposis, and a may well be an underestimate of the true proportion of individuals, among those dying before age 40 from primary cancer of the large bowel, whose cancer is secondary to polyposis. Therefore the true value of f is probably higher than the calculated value. It should be noted that because persons with the gene for polyposis have a decreased life expectancy, the frequency of such persons in the general population will be less than (approximately two-thirds) that at birth.

RELATIVE FITNESS OF INDIVIDUALS WITH MULTIPLE POLYPOSIS

Fitness, in a population sense, is measured by ability to produce children and is a function of viability (from birth through the reproductive period) and fertility (in the narrow sense, once the reproductive period is reached). The fitness of a class of individuals relative to that of another class may, under ideal conditions, be measured by the ratio of the expectation *at birth* of the number of live-born children to be produced by a live-born individual of the first class, to the corresponding expectation of the second class. If the first class is composed of heterozygotes for a rare dominant gene which lowers fitness, such as that for multiple polyposis, and the second class is the remainder, all normal in this respect, of a population in equilibrium, this estimate of relative fitness (W) is also the proportion of the dominant genes transmitted from one generation to another. An estimate of W for multiple polyposis will measure the degree of natural selection for or against bearers of the gene and is required in the indirect estimation of the mutation rate. Two independent methods of calculating W are available.

The direct calculation of relative fitness from the observed reproductive performance of affected persons and their normal sibs is complicated by the late manifestations of multiple polyposis in some individuals. Depending on the method of ascertainment of the data, a possible further complication is the tendency of members of large families to have more children than members of small families (Fisher, 1930). The first difficulty is somewhat reduced by using only sibships whose apparently normal members were all over 40 years of age at the time of investigation or at death. Six such sibships are available; they are described in Table 3. It is clear that appearance of polyposis in persons after 40 can make this estimate of W an underestimate since, on the average, these persons are very probably more fertile than persons affected before 41. In theory, this restriction introduces a possible bias from deaths before age 41 of individuals lacking the gene for polyposis. In our sample, however, there were no such deaths among the apparently normal members of these sibships, thus eliminating this source of bias. Unfortunately, the 6 sibships are heterogeneous with regard to ascertainment, 2 containing a propositus, 2 containing one parent of a propositus, 1 containing two parents of propositi, and 1 containing first cousins of the propositus. Lacking a suitable weighting procedure to correct for ascertainment and sibship size frequency, it seems best to calculate W simply as the ratio of the mean number of children from persons affected with polyposis to the corresponding mean for the apparently normal sibs, using pooled data of the six sibships. From the totals of Table 3, the mean for affected persons is $45/18 = 2.50$ and that for "normal" persons is $78/23 = 3.39$, giving an estimate of W of 0.74. If, in fact, there is no

TABLE 3.—DESCRIPTION OF SIBSHIPS USED IN DIRECT CALCULATION OF RELATIVE FITNESS
All "normal" individuals were over age 40 at time of investigation or at death. See "Remarks" for details on omission of certain individuals.

Sibship	Size	Affected		"Normal"		Remarks
		(Usable) number	No. of children	(Usable) number	No. of children	
1554 (II, 6–9)	4	3	8	1	2	All persons usable; first cousins of the propositus.
809 (II, 1–5)	5	2	3	2	4	II-1, parent of propositus, excluded.
1554 (I, 1–7)	7	2	6	4	10	I-7, parent of propositus, excluded.
1826 (II, 1–8)	8	2	9*	4	12	II-1 and II-5, parents of propositi, excluded.
1826 (III, 20–30)	11	5†	16	4	26	III-24 and III-26 excluded because of uncertain status with regard to polyposis. A propositus is included.
1862 (II, 1–12)	12	4‡	3	8	24	All persons usable; the propositus is included.
Total.....	—	18	45	23	78	

* Twins counted as one individual.
† III-27 counted as affected.
‡ II-4 counted as affected.

difference in the means, the probability of obtaining an estimate as low as or lower (single-tail test) than the one observed is 0.05.

The second method for estimating relative fitness is indirect and requires the assumptions that (1) the only effect on W of the gene for multiple polyposis is through the death, from cancer secondary to polyposis, before the end of the reproductive period, of some affected individuals, and (2) that until their death persons with the gene reproduce at the same rate as persons lacking the gene. There appear to be no data suggesting that the first assumption is incorrect with regard to the *biological* action of the gene, although the possibility of pleiotropic effects must always be considered. The second assumption, judging from the present data, appears to be reasonable for the period up to the time of first bowel complaint. The possibility exists that some children of affected parents will restrict their reproduction when they are concerned about the appearance of polyposis in their own descendants. Such restriction was not apparent in our data. Reproductive capacity is, of course, impaired in the interval between onset of symptoms and death. The bias resulting from making the second assumption, however, appears to be only several percent. This bias is discussed below.

This estimate of W, then, will actually be a function of relative survival to and through the reproductive period. The quantity, which we here equate with W, we may term the "relative reproductive span" or RRS.[3]

[3] Although the use of RRS as a measure of fitness appears permissible in the present context, it should be recognized that for many dominant traits this is not the case. Thus, Crowe, Schull, and Neel (in press) find that although part of the effect of the gene responsible for neurofibromatosis on

If

p_x = the proportion of births, among all births, which occur to parents of age x (mean of values of paternal and maternal age distributions),

l_x = the proportion of all live-born individuals with the gene for polyposis who survive to age x,

L_x = the proportion of all live-born individuals lacking the gene for polyposis who survive to age x, and

d_i = the proportion of deaths, among all deaths from cancer secondary to polyposis, which occur at age i,

then

$$RRS = \sum_x p_x \left(\frac{l_x}{L_x}\right) \sim \sum_x p_x \left(1 - \sum_0^{x-1} d_i\right),$$

summation extending over the longest life span. These relations are derived in the appendix.

To obtain estimates of d_i both the data of the present study and the excellent study of Dukes (1952) were used. Utilizing only individuals dying at a known age from medically diagnosed cancer of the large bowel and who either were medically diagnosed as having multiple polyposis or who were close relatives of persons so diagnosed, 29 ages at death were obtained from the present study and 62 from that of Dukes. The period over which these deaths occurred was 1916–1953, with a mean at 1940.4 for the present study, and 1882–1951, with a mean at 1930.2 for the 59 deaths of Dukes' study for which data are given. There was one other death before 1900 among these 59. The distribution of these ages is given in Table 4 and the means, standard deviations, and standard errors in Table 5. It is seen that the means and variances of males and females do not differ significantly within the two studies nor do the means and variances between the studies. It was therefore considered appropriate to combine all data, resulting in a mean age of death of 40.21 ± 1.23 years and a standard deviation of 11.77 years. In spite of a marked dip at the 35–39 year interval the combined distribution does not differ significantly from that expected of a normal curve with the same mean and variance. To reduce the effects of chance fluctuation in the proportions of deaths in the 5-year-age intervals, it seemed advisable to estimate d_i from the normal curve.[4] Values of d_i through the reproductive period are

fitness is exerted through the early death of a few individuals with the trait, the major effect is a decreased marriage rate on the part of persons with the trait, as well as impaired fertility after marriage. At the other extreme, Panse (1942) and Reed and Palm (1951) have suggested that despite the early death of some persons with Huntington's chorea, the net fertility of affected persons is actually greater than normal.

[4] It is recognized that it is unlikely that the distribution is normal. More extensive data would doubtless make this apparent and also would permit a decision as to whether or not the two observed peaks in Table 5, at 30–34 years and 45–49 years, are real. At the present stage of our knowledge the assumption of normality, for purposes of calculation, seems justified by the symmetry of the mean with respect to the extremes and by the similarity of the distributions around the modes of age-at-death curves for various diseases. The mean age at death in polyposis of 40 years, with extremes at about 10 and 70 years, suggests a symmetrical distribution, while the bell-shaped distributions around the mode found in many diseases, including various cancers, suggest that the normal distribution is not unreasonable.

TABLE 4.—DISTRIBUTION OF AGES AT DEATH OF PERSONS DYING FROM CANCER OF THE COLON OR RECTUM ARISING FROM MULTIPLE POLYPOSIS (CANCER MEDICALLY DIAGNOSED)
Present study and Dukes, 1952

Age at death	Present study*	Dukes, 1952†	Total
0–4	0	0	0
5–9	1	0	1
10–14	0	0	0
15–19	2	0	2
20–24	0	3	3
25–29	3	6	9
30–34	7	14	21
35–39	1	8	9
40–44	4	8	12
45–49	6	9	15
50–54	3	5	8
55–59	0	6	6
60–64	1	2	3
65–69	0	1	1
70–74	1	0	1
75+	0	0	0
Total	29	62	91

* Omitting the 4 persons selected for having died under 40 years of age. Multiple polyposis either medically diagnosed or inferred from existence of a parent or child with medically diagnosed multiple polyposis.

† All individuals are members of kindreds containing at least one medically diagnosed case of multiple polyposis. Diagnosis of cancer established by one of the following: medical or death certificate, medical examination, hospital record.

given in Table 6. Values of p_x for Michigan are available from the vital statistics of the U. S. Bureau of the Census for 1934 and later years. In order to make the data for p_x and d_i more comparable, p_x for Michigan in 1935 was used. The values of p_x and the calculation of the relative reproductive span are given in Table 6. This estimate is 0.78, slightly higher than the previous estimate of W of 0.74 obtained from the observed reproductive performance of 6 sibships. This latter estimate is considered less reliable than the RRS estimate because (1) it is based on few data, (2) the diagnoses of a number of individuals are not certain, and (3) the method to correct for ascertainment and sibship size is not apparent.

Two known biases of this estimate should be considered. One bias is in the formula for the RRS. The derivation of the formula used shows that it slightly underestimates the true value of l_x/L_x, and hence of RRS. A bias in the opposite direction results from the assumption that reproduction continues until death from cancer secondary to polyposis. Among the 29 cases of the present study used in estimating the mean age at death, data were available in 17 cases on the interval between onset of significant bowel complaints and death from cancer. This interval varied from less than one year to 14 years, with a mean of 3.2 years. Of these 17 cases, 10 died under age 40 and the intervals varied from less than one year to 8 years, with a mean at 2.8 years. Since, on the average, reproductive capacity does not end with onset of bowel com-

TABLE 5.—AGE AT DEATH FROM CANCER OF THE COLON OR RECTUM OF PERSONS WHOSE CANCER AROSE FROM MULTIPLE POLYPOSIS (CANCER MEDICALLY DIAGNOSED)
Present study and Dukes, 1952

Source*	Individuals	Number		All deaths		
		Dying < 40 years	All deaths	Mean	Standard deviation	Standard error
Present study	Males	6	15	42.00	13.76	3.55
	Females	8	14	35.57	11.54	3.08
	Males and females	14	29	38.90	12.93	2.40
Dukes, 1952	Males	18	34	40.24	10.59	1.82
	Females	13	28	41.54	12.17	2.30
	Males and females	31	62	40.82	11.25	1.43
Present study and Dukes, 1952	Males and females	45	91	40.21	11.77	1.23

* See Table 4 for further description.

TABLE 6.—CALCULATION OF THE RELATIVE REPRODUCTIVE SPAN (RRS). SEE TEXT FOR SYMBOLS

x, i (Age interval)	p_x	d_i	$1 - \sum_0^{x-1} d_i$ *
0–14	.000	.016	.992
15–19	.063	.027	.970
20–24	.250	.0555	.929
25–29	.275	.0935	.855
30–34	.200	.138	.739
35–39	.125	.162	.589
40–44	.060	.167	.425
45–49	.019	.138	.272
50–54	.006	.099	.153
55+	.002	.104	.052
Total	1.000	1.000	—

* $\sum_0^{x-1} d_i$ as used here represents the proportion of deaths up to the mid-value of each age interval.

$$RRS = \sum_0^{100} p_x \left(1 - \sum_0^{x-1} d_i\right) = 0.78$$

plaints but will steadily decline from that time, the above assumption involves an error of a year or so. A method based on this assumption will overestimate the true value by a few percent. It is not possible to say whether these two biases will cancel out. The relative fitness at the present may be higher than the calculated value since this value depends on deaths with a mean around 1935. Better diagnosis and greater use of radical surgical procedures in the future treatment of multiple polyposis may be expected to increase the life expectancy of affected individuals and so raise the relative fitness.

CALCULATION OF THE MUTATION RATE ESTIMATE

Because of the late appearance of polyps, and symptoms subsequent to polyps, in some individuals (e.g., Dukes (1952) studied an individual who at 38 was normal to sigmoidoscopy but at 44 had polyps and at 56 developed carcinoma of the colon; in the present study III-5 in Kindred 1826 first had intestinal complaints at age 57; polyposis was diagnosed at 58), it is not possible, at present, to be certain that any individual will not later develop polyposis, although it seems quite unlikely that polyps will first appear after age 50. Therefore, it is not feasible to calculate a direct estimate of the mutation rate. Only one kindred (1963) of the present study presents a reasonable case for mutation. The propositus of this kindred has 9 sibs whose ages range from 25 to 40 and both parents are living at age 69; all are reported free of significant intestinal complaint. Several other kindreds may demonstrate mutation but again no proof can be offered. At present we are forced to rely on an indirect calculation of the mutation rate.

If the mutation rate/gene/generation is m and the population is in equilibrium with respect to production and loss of genes for polyposis, the following customary equation applies:

$$m = \frac{f(1 - W)}{2}.$$

The estimate of W from the relative reproductive span is more reliable and will be used here. Substitution of the values for f and W gives

$$m = \frac{1.21 \times 10^{-4}(1 - .78)}{2} = 1.3 \times 10^{-5}$$

Since f is probably an underestimate, the value of m may, perhaps, actually be up to twice this value. The probable bias of W is not known. Considering the bias of f, it seems unlikely that the true value of W would be such that m would be less than about three-fourths of the calculated value. m probably lies within the range $1\text{–}3 \times 10^{-5}$.

DISCUSSION

The present study is one of a series of investigations on mutation rates carried out by this Clinic. Some of the problems inherent in such investigations, as they have impressed themselves on us, have been discussed elsewhere (Neel, 1952; Neel and Schull, 1954; see also Haldane, 1949, Nachtsheim, 1954; and Vogel, 1954). In addition to certain methodological questions which are common to all mutation rate studies, each trait selected for study has raised particular difficulties more or less unique to that trait. Thus, in the case of multiple polyposis, we are confronted with the fact that it is very difficult to demonstrate that any particular "sporadic" case is due to mutation. As a consequence, no use can be made of the "direct" method of estimating mutation rate (i.e., from the observed frequency of sporadic cases). Both the necessity of performing sigmoidoscopic and X-ray studies on the parents of affected persons and the late appearance of polyps and symptoms in some individuals

preclude even an approximate direct determination of mutation rate. In passing, we may note that although we recognized the unpleasant nature of these diagnostic studies, we were unprepared for the lack of cooperation sometimes encountered, especially since the studies were so obviously to the advantage of the person being investigated.

Since the direct method was not feasible, we have employed the indirect approach, based on estimates of frequency and relative fitness. The estimation of the relative fitness of affected individuals proved to be difficult, partly because it approaches that of the general population. The observed relative fitness of about 0.8 is appreciably higher than that of most other dominant traits for which mutation rate estimates exist. The most critical assumption underlying the use of the indirect method, in this and other studies, is that genetic equilibrium obtains, i.e., that the loss of genes (for the trait in question) in each generation is balanced by the appearance of new genes through mutation. While there are on record a number of kindreds highly suggestive of the occurrence of mutation with respect to the gene for multiple polyposis (e.g., our Kindred 1963; families 12 and 21, Dukes, 1952; 1 kindred, Gardner and Woolf, 1952), the assumption that the polyposis locus is in genetic equilibrium because of mutation from the normal allele to the gene for multiple polyposis is less tenable here than in the case of some other dominant traits which have been utilized in mutation rate studies. On the other hand, the proportion of propositi whose disease is not clearly inherited in both this study and that of Dukes (1952), while obviously an overestimate of the true proportion of sporadic cases, is in keeping with the hypothesis that each generation about one-quarter of the polyposis genes must arise through mutation if genetic equilibrium exists.

A few general remarks concerning the philosophy of this Clinic with regard to mutation rate studies are perhaps appropriate at this point. Almost every mutation rate estimate advanced to date—not excluding our own—can be subjected to severe criticism. It seems not only possible but probable that many of the existing estimates err by a factor of two or even more. At this stage in our developing appreciation of the problem, this does not seem to us a serious deterrent to such studies. The present challenge is to fix the order of magnitude of the phenomenon in a relatively long-lived animal, by a series of studies on as many traits as possible. Later, as techniques improve and the outlines of the problem become clearer, greater accuracy will be possible, as will a comparison of human mutation rates with those of other forms. The present estimate of $1\text{–}3 \times 10^{-5}$ mutations/gene/generation falls well within the range of other available estimates, all of which, of course, assume that only a single locus is involved. The dangers inherent in attempting to generalize at the present time from this and the other existing estimates to *all* genes have been discussed elsewhere (Neel and Schull, 1954). On the other hand, each new estimate strengthens the foundation of fact from which generalizations may someday be possible.

CONCLUSIONS AND SUMMARY

In a study of the genetics of multiple polyposis of the colon, a rare dominant trait, special emphasis has been given to the estimation of the frequency and relative fitness of individuals bearing the gene, and of the mutation rate of the gene. The

material of this study consists of 23 kindreds, including 70 certainly and 13 probably affected persons. Fourteen of the 23 kindreds contain two or more affected members. The mean age at death from cancer of the colon or rectum subsequent to polyposis, in 91 very probably or certainly affected individuals in the present study and in that of Dukes (1952), was 40.21 ± 1.23 years. From a survey of the University of Michigan Hospital records between 1935 and 1944, the proportion of individuals, among persons dying before age 40 from cancer of the colon or rectum who also had multiple polyposis, was estimated. This estimate, which is minimal, is 0.103 ± 0.040. From these facts and the known distribution of age at death from cancer of the colon and rectum in Michigan, an estimate of the minimum frequency at birth of individuals with the gene for multiple polyposis was obtained: $(1.21 \pm 0.49) \times 10^{-4}$ or 1 in 8,300.

Two estimates of the relative fitness of individuals with the gene have been derived, the more reliable being that of the "relative reproductive span." This is a weighted measure of the survival, to and through the reproductive period, of persons with the gene relative to that of persons lacking the gene. This estimate is 0.78. Using these estimates for frequency and relative fitness, and considering the known biases involved, the mutation rate is estimated to be $1\text{–}3 \times 10^{-5}$/gene/generation.

An appendix derives equations for the estimation of relative reproductive span.

REFERENCES

Bacon, H. E. 1949. *Anus, Rectum, and Sigmoid Colon: Diagnosis and Treatment.* 3rd ed. Philadelphia: J. B. Lippincott. 2 vols. 1127 pp.

Cockayne, E. A. 1927. Heredity in relation to cancer. *Cancer Rev.*, 2: 337–347.

Cripps, H. 1882. Two cases of disseminated polypus of the colon. *Tr. Path. Soc., London*, 33: 165–168.

Crowe, F. W., Schull, W. J., and Neel, J. V. Multiple Neurofibromatosis, *American Lectures in Dermatology* series. Springfield: C. C. Thomas. (In press.)

Dublin, L. I., Lotka, A. J., and Spiegelman, M. 1949. *Length of Life.* New York: The Ronald Press Co. 379 pp.

Dukes, C. E. 1952. Familial intestinal polyposis. *Ann. Eug.*, 17: 1–29.

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press. 272 pp.

Gardner, E. J., and Richards, R. C. 1953. Multiple cutaneous and subcutaneous lesions occurring simultaneously with hereditary polyposis and osteomatosis. *Am. J. Human Genet.*, 5: 139–147.

Gardner, E. J., and Woolf, C. M. 1952. Intestinal polyposis and carcinoma originating from a mutation in a family group. *Cancer*, 5: 695–699.

Gianturco, C., and Miller, G. A. 1953. Program for the detection of colonic and rectal polyps. *J.A.M.A.*, 153: 1429–1430.

Haldane, J. B. S. 1949. The rate of mutation of human genes. Proc. 8th Int. Cong. Gen., *Hereditas*, Supp. Vol.: 267–273.

Helwig, E. B. 1947. The evolution of adenomas of the large intestine and their relation to carcinoma. *Surg. Gynec. and Obst.*, 84: 36–49.

Jeghers, H., McKusick, V. A., and Katz, K. H. 1949. Generalized intestinal polyposis and melanin spots of the oral mucosa, lips, and digits; a syndrome of diagnostic significance. *New England J. Med.*, 241: 993–1005.

Nachtsheim, H. 1954. Die Mutationsrate menschlicher Gene. *Naturwiss.*, 17: 385–392.

Neel, J. V. 1952. The study of human mutation rates. *Am. Nat.*, 86: 129–144.

Neel, J. V., Bolt, R. J., and Pollard, H. M. 1954. A pedigree of multiple polyposis of the colon. *Gastroenterology*, 26: 1–11.

Neel, J. V., and Schull, W. J. 1954. *Human Heredity.* Chicago: University of Chicago Press. 361 pp.

Panse, F. 1942. *Die Erbchorea, eine klinisch-genetische Studie.* Leipzig: Georg Thieme.

Reed, S. C., and Palm, J. D. 1951. Social fitness versus reproductive fitness. *Science*, 113: 294–296.

SWINTON, N. W., AND HAUG, A. D. 1947. The frequency of precancerous lesions in the rectum and colon. *Lahey Clin. Bull.*, 5: 84–88.

VOGEL, F. 1954. Über Genetik und Mutationsrate des Retinoblastoms (Glioma retinae). *Z. menschl. Vererb.-u. Konstitutionslehre*, 32: 308–336.

APPENDIX

DERIVATION OF EQUATIONS FOR ESTIMATING THE RELATIVE REPRODUCTIVE SPAN (*RRS*)

The direct estimation of the relative fitness of individuals with a disease like multiple polyposis, from their observed reproductive performance, may be unsatisfactory because of the late onset of the disease in some individuals. In such a situation an indirect estimate may be necessary. If reliable data are available on the age at death due to the disease and the assumption is justified that early death from the disease is the only factor lowering the relative fitness of such individuals, it is possible to obtain an indirect estimate which may be more reliable than the direct.

Consider two cohorts, C_1 and C_2, each containing N new-born individuals (N being large). The cohorts differ only because all members of C_1 have the gene for polyposis and all members of C_2 lack it. If the cohorts are enumerated each year from birth ($x = 0$) to death, and births to cohort members at each year x are noted, we may define the following terms:

$$f_x = \text{age-specific fertility} = \frac{\text{number of births to a cohort in the year } x}{\text{number of individuals of the cohort alive at the beginning of year } x},$$

l_x = proportion of C_1 surviving to age x,

L_x = proportion of C_2 surviving to age x.

f_x is postulated to be equal for the two cohorts, i.e., as long as an individual with the gene is living he is assumed to be as fertile as individuals lacking the gene. (The small bias introduced by this assumption is discussed in the text.)

Our observed data on births and deaths are not in terms of cohorts but are obtained from a population composed of all ages. It is therefore pertinent to note that for a normal population in equilibrium the age-distribution will be that of a "life table population" (Dublin, Lotka, and Spiegelman, 1949), i.e., L_x is the proportion of persons who are age x. We shall make use of this equivalence below. In terms of the observed population data we may define two further terms:

p_x = the proportion of births, among all births, which occur to parents of age x (mean of values of paternal and maternal age distributions),

d_i = the proportion of deaths, among all deaths, from cancer secondary to polyposis which occur at age i.

The mean number of children (live-born) ever born to the N members of C_1 is

$\sum_x l_x f_x$, the summation extending to the longest life span. The corresponding mean for the N members of C_2 is $\sum_x L_x f_x$. Therefore,

$$W = \text{relative fitness} = \frac{\sum_x l_x f_x}{\sum_x L_x f_x}.$$

But, for any age j

$$f_j = \frac{N p_j \sum_x L_x f_x}{N L_j}, = \frac{p_j \sum_x L_x f_x}{L_j},$$

so that

$$W = \sum_x p_x \left(\frac{l_x}{L_x}\right),$$

under the above assumptions, summation again extending to the longest life span.

If the assumption stated above, i.e., that early death is the only effect of the gene on fitness, does not hold, $\sum_x p_x \left(\frac{l_x}{L_x}\right)$ will not equal W. We may define $\sum_x p_x \left(\frac{l_x}{L_x}\right)$ as the *relative reproductive span* (RRS). Its usefulness for the present study as a means of estimating W, requires the above assumption, but, for other traits where a direct estimate of W is available, it may be helpful in testing the validity of this assumption.

Let

q_x = probability that a person with the gene for polyposis who reaches his xth birthday will die from cancer secondary to polyposis before his $x + 1$th birthday,

r_x = probability that a person lacking the gene for polyposis who reaches his xth birthday will die before his $x + 1$th birthday.

Since persons with the gene for polyposis, before onset of symptoms, are assumed not to differ from persons without the gene, and, if after onset of symptoms they are considered to be still subject to all the other causes of death from which persons without the gene die, the probability that such a person will die within the year following his xth birthday is $q_x + r_x - q_x r_x$ (very nearly; shorter time intervals would make this more exact). Since $q_x r_x$ for ages of interest to us, i.e., up to the end of the reproductive period, is small compared to q_x and r_x, we may neglect this term and write

$$l_x = 1 - \sum_0^{x-1} l_i(q_i + r_i)$$

and

$$L_x = 1 - \sum_0^{x-1} L_i r_i.$$

Neglecting products of the form $q_i r_j$, etc. we may write

$$L_1 = l_1 + l_0 q_0 ,$$

$$L_2 = l_2 + l_1 q_1 + l_0 q_0 ,$$

and, in general,

$$L_x = l_x + \sum_0^{x-1} l_i q_i .$$

Therefore

$$L_i r_i = l_i r_i + r_i \sum_0^{i-1} l_j q_j$$

and

$$\frac{l_x}{L_x} = \frac{1 - \sum_0^{x-1} l_i (q_i + r_i)}{1 - \sum_0^{x-1} l_i r_i - \sum_0^{x-1} \left[r_i \sum_0^{i-1} l_j q_i \right]}.$$

Let

$$K_x = \sum_0^{x-1} \left[r_i \sum_0^{i-1} l_j q_i \right].$$

Then

$$\frac{l_x}{L_x} = 1 - \frac{\sum_0^{x-1} l_i q_i - K_x}{1 - \sum_0^{x-1} l_i r_i - K_x}$$

Since, nearly,

$$\sum_0^{100} l_i q_i + \sum_0^{x-1} l_i r_i + \sum_x^{100} l_i r_i = 1,$$

and putting

$$\sum_0^{100} l_i q_i = P$$

where P is the proportion of persons born with the gene for polyposis who die from cancer secondary to polyposis, and noting that, by definition,

$$\sum_0^{x-1} d_i = \frac{\sum_0^{x-1} l_i q_i}{\sum_0^{100} l_i q_i} = \frac{\sum_0^{x-1} l_i q_i}{P},$$

then

$$\frac{l_x}{L_x} = 1 - \frac{\sum_0^{x-1} d_i - \frac{K_x}{P}}{1 + \frac{\sum_x^{100} l_i r_i}{P} - \frac{K_x}{P}}.$$

Since only $\sum_0^{x-1} d_i$ is known, it is necessary to consider the appropriateness of estimating l_x/L_x by $1 - \sum_0^{x-1} d_i$. The magnitude of P is not known with certainty but study of kindreds in which the gene for polyposis is segregating yields information since we find (in this study and in Dukes, 1952) that the observed proportion of cases of polyposis among adult offspring of an affected parent approaches the expected 0.5 and that large bowel cancer usually follows before age 50 is reached, although there are several instances of later onset. It seems unlikely that during the present century any large proportion of individuals born with the gene fails to manifest multiple polyps and subsequent cancer. Such cancer, until recently, must usually have proved fatal. The 1940 American life table indicates that about 83 per cent of live-born individuals will be alive at age 50 and, therefore, persons with the gene are likely to survive to the age where cancer from malignant degeneration of polyps is prevalent. With these considerations, the value of P would be expected to be at least of the order of 70 per cent. For low values of x, e.g., in the case of multiple polyposis under 25 years of age, the expression for l_x/L_x is seen to reduce to $1 - \frac{\sum_0^{x-1} d_i}{1 + \frac{1-P}{P}}$, so that if $\sum_0^{x-1} d_i = 0.1$ and $P = 0.7$, $l_x/L_x = 0.93$, while the approximation $1 - \sum_0^{x-1} d_i$ gives 0.90. A higher value for P, of course, makes this approximation better. The approximation should be of this order in the middle-aged range and, for high x, should be better yet.

If we introduce this approximation to l_x/L_x into our previous formula for *RRS*, we have

$$RRS = \sum_x p_x \left(1 - \sum_0^{x-1} d_i\right),$$

the relation to be derived. This estimate should slightly underestimate the true value of *RRS*.

Aspects of Genetics in Psychology

ARNE TRANKELL

Department of Psychology, University of Göteborg

INTRODUCTION

THE STUDY OF heredity in psychology has mainly concentrated on twins with the specific objectives conditioned by this kind of material. Only sporadical attempts have been made to study heredity according to Mendelian principles. The reason for this is probably to be found in the fact that most behavior patterns are not stable enough to permit a Mendelian hypothesis. The question may be raised, however, if psycho-genetics does not have at its command an instrument which could be used far more often than has been the case. This instrument is population genetics.

In his book, "Mathematical Methods for Population Genetics", Dahlberg (1947), has given formulae which have applicability on a number of problems in human genetics. The working conditions encountered by psychology, especially as regards environmental influences and methods of measurement, necessitates, however, the construction of special formulae to meet its demand. Formulae not based on Mendelian principles are, of course, out of the question. The suggestions presented in this paper are therefore nothing but applications of the Mendelian principles on problems of particular interest to the psychologist.

In order to be able to draw any conclusions regarding the course of inheritance of a specific trait it is necessary to know the proportions of the phenotypes that may occur. This may be possible when panmixia prevails which is probably the case for a number of mental traits with a hereditary background. The simplest case is found when the inheritance is monofactorial and there are no more than two alleles. This paper mainly deals with this case.

When it comes to mental traits, possible deviations from the classical Mendelian principles of manifestation, must be taken into account. The reason is that psychology has to work with behavior and not with morphological traits, and the behavior is less constant and more easily influenced by environmental factors, than other hereditarily conditioned expressions. It is probably more of a rule than an exception that behavior tendencies are so influenced by the environment that the original effect of the genotype no longer can be observed. If we conceive, for instance, that a certain behavior pattern is promoted by a dominant gene and that the proportion between the alleles is such as to let the majority be characterized by the dominant behavior, a social pressure may arise in the population promoting the dominant behavior at the expense of the recessive behavior. The recessive homozygotes are thereby forced to adopt the more sanctioned dominant behavior. An example of this is found in the strive of the lefthanded to accept and adopt the right-handed behavior which is exrpessed in a decreasing frequency of left-handedness with increasing age.

Received May 22, 1955.

Another difficulty to be mastered by psycho-genetics lies in the technique of measurement. The recording of behavior tendencies is more difficult than any other. This is due in main to the fact that it is nearly impossible to obtain generally accepted criteria, but secondly also because of the inconsistency of behavior in itself. We have therefore to take into account that because of measurement-technical difficulties, every recording of trait-carriers will end up more or less incomplete. This indicates, that, in what is technically referred to as the "incomplete penetrance" of the gene, we have also to consider a certain decrease in manifestation because of incomplete recording of the trait-carriers.

A THEORETICAL MODEL

Many traits with a hereditary background are influenced by a directed reduction in manifestation, which means that the environment and the incomplete recording are both active in a particular direction, as in handedness from the left-handed toward the right-handed behavior. When this is the case, we may mathematically eliminate the effect of the incomplete manifestation by including in the calculation the deviations in the proportions of genotypes within different groups caused by the incomplete penetrance. In some cases it will not only be possible to test the genetic hypothesis, but also to calculate the proportion of the alleles in the studied population.

One way to perform this task will now be described for the case when the theoretical model assumes complete penetrance of the dominant trait in the heterozygotes and the dominant homozygotes, but only partial penetrance of the recessive trait in the recessive homozygotes.

As a point of departure we assume that there are two alleles designated D and R. Their relative proportions are designated d and r ($d + r = 1$). We further assume that panmixia prevails for the genotypes in question. According to the Hardy-Weinberg law the proportions of the different genotypes are d^2 DD, 2dr DR and r^2 RR, where $d^2 + 2dr + r^2 = 1$. Connected with these genotypes we assume two alternative phenotypes, non trait-carriers and trait-carriers. The trait is recessive and connected with the allele R. It is only partially manifested by the recessive homozygotes.

The studied population (a number of families of the main population chosen at random) is divided into two generations: P-generation (the parents) and F-generation (the children). Since we have to consider different manifestations at different age-levels and the recording instruments may be differently effective at different age-levels, the frequency of individuals recorded as trait-carriers in the P-generation is designated a, and the same proportion in the F-generation is designated b.

The F-generation is divided into three sub-groups according to the recorded phenotype of the parents. Group I contains children whose parents both have been recorded as trait-carriers; group II children belonging to families where one of the parents has been recorded as a trait-carrier, while group III includes children of families where none of the parents has been recorded as a trait-carrier.

According to the given definitions the proportions of the various genotypes in the

P-generation are (the recessive homozygotes manifesting the recessive trait have been given index $_r$ and the rest of the recessive homozygotes index $_d$):

Recorded phenotype	Genotype	Proportion
Non trait-carriers	DD	d^2
	DR	2dr
	RR_d	$r^2 - a$
Trait-carriers	RR_r	a

In family-type I the genotypes of the children are determined by the following type of mating: $(RR_r)(RR_r)$. All individuals in the F-generation will therefore be of genotype (RR) and the proportion of recessive homozygotes among the descendants is consequently:

$$\frac{a^2}{a^2} = 1.$$

In family-type II the progeny is determined by the combination $2(DD + DR + RR_d)$ (RR_r). The porportion of recessive homozygotes among the descendants is consequently:

$$\frac{2dra + 2a(r^2 - a)}{2a(1 - a)} = \frac{r - a}{1 - a}.$$

In family-type III the progeny is determined by the combination $(DD + DR + RR_d)(DD + DR + RR_d)$. The proportion of recessive homozygotes among the descendants is consequently:

$$\frac{d^2r^2 + 2dr(r^2 - a) + (r^2 - a)^2}{(1 - a)^2} = \frac{(r - a)^2}{(1 - a)^2}.$$

With knowledge of the proportions of (RR) in the three subgroups of the F-generation it is possible to formulate expressions for the recorded number of individuals who manifest the recessive trait within each group. The following designations are then used. The total number of children within the subgroups are designated N_1, N_2, and N_3, and the recorded number of children who manifest the recessive trait are designated X_1, X_2, and X_3. The coefficient b/r^2 indicates the portion of recessive homozygotes in the F-generation recorded as trait-carriers under the prevailing penetrance conditions, since b is the fraction of recorded trait-carriers in the F-generation, while r^2 is the proportion of all recessive homozygotes. We then get the following equations:

$$\text{Family-type I} \qquad \frac{b}{r^2} \cdot N_1 = X_1 \tag{1}$$

$$\text{Family-type II} \qquad \frac{b}{r^2} \cdot \frac{r - a}{1 - a} \cdot N_2 = X_2 \tag{2}$$

$$\text{Family-type III} \qquad \frac{b}{r^2} \cdot \frac{(r-a)^2}{(1-a)^2} \cdot N_3 = X_3 \qquad (3)$$

In applying the formulae only r-values consistent with the original hypothesis are accepted. When the penetrance of the recessive homozygotes is at its maximum value e.g., when all recessive homozygotes are recorded as trait-carriers, a is equal to r^2. This gives us the theoretical minimum value of r, namely $r_{min} = \sqrt{a}$ (if b is greater than a, $r_{min} = \sqrt{b}$). Values of r which are smaller than the minimum value cannot be accepted.

When we are in possession of a set of empirical data, the Mendelian hypothesis may be tested by X^2-analyses of the deviations between actual values and the expected amount of recorded trait-carriers in the sub-groups corresponding to different r-values. The analyses can be restricted to r-values around an approximately determined value, which may be arrived at by inserting into the equations the known values of N, X, a and b and solving for r. The three determinations of r which are then obtained, will be approximately equal when the hypothesis is unrejectable. In this case we will always find a range of r, within which each r-value describes the variation of frequencies of trait-carriers in the sub-groups. The r value giving the smallest X^2 may be taken as an estimation of the proportion of R alleles in the population in question. This estimation is independent of the extent of the penetrance. The determination of the proportions of the alleles is therefore independent of the criteria used in recording the trait-carriers.

INHERITANCE OF HANDEDNESS

1. Hypothesis

The procedure will be exemplified through analyses of three studies of the inheritance of left-handedness. They have all been carried out in USA and are the only internationally known studies of left-handedness permitting an application of population genetics. They have been performed at different times and are distinguished by entirely different criteria of left-handedness. Much varying frequencies of left-handedness have therefore been obtained. The analyses show, however, that the three populations have the same genetic constitution.

If as a point of departure we assume that right-handedness is conditioned by the dominant gene in a pair of alleles, it will hold true that left-handedness appears only in the absence of the dominant gene. Because of the social pressure promoting right-handed behavior, we have to calculate with the possibility that a portion of the recessive homozygotes so strongly suppress the recessive behavior tendencies that it is impossible to distinguish them from right-handers being right-handed due to the dominant gene. The more incomplete the recording of left-handed tendencies is, the smaller is the fraction of recessive homozygotes recorded as trait-carriers. It may as well be assumed that the absence of the dominant gene activates other genes which control the strength of the left-handed tendencies. It is also possible that the weaker forms of left-handedness cannot be distinguished from the right-handed behavior determined by the action of the dominant gene.

2. Rife's material

A wellknown investigation of the inheritance of left-handedness is made by D. C. Rife (1940). It is made up of two parts: one twin-study implying the presence of genetically determined differences in degrees of left-handed tendencies, and a population-genetic material including 687 families collected by Rife among the students at Ohio State University. On the basis of this latter material Rife concluded that "left-handers are more likely to have left-handed children than are the right-handers". The criterion of left-handedness used by Rife was 10 selected acts known to be carried out by most people with their right hand. Individuals indicating the use of their left hand in one or more of these 10 acts were recorded as left-handed. The P-generation contained 1374 individuals, 72 of which were recorded as left-handed (a = 0.0524).

The F-generation contained 2178 individuals, 191 of which were recorded as left-handed (b = 0.0877).

Family-type I (both parents left-handed) contained 11 children, 6 of which were recorded as left-handers. An approximate r-value is determined by the equation $\frac{0.0877}{r^2} \cdot 11 = 6$, which gives $r = 0.401$.

Family-type II (one parent left-handed) contained 174 children, 34 of which were recorded as left-handers. An approximate r-value is determined by the equation $\frac{0.0877}{r^2} \cdot \frac{r - 0.0524}{0.9476} \cdot 174 = 34$, which gives $r = 0.414$.

Family-type III (both parents right-handed) contained 1993 children, 151 of which were recorded as left-handers. An approximate r-value is determined by the equation $\frac{0.0877}{r^2} \cdot \frac{(r - 0.0524)^2}{0.9476^2} \cdot 1993 = 151$, which gives $r = 0.439$.

The three determinations of r correspond well. The X^2-analyses show that r-values within the range of 0.51–0.35 are acceptable on the 5 per cent level, but values outside these limits make the hypothesis unacceptable. The smallest X^2-sum is obtained, when r equals 0.410, a value which may be looked upon as an estimation of the proportion of R in the population, from which Rife's material was collected. The deviations between the actual values and the values corresponding to various r-hypotheses are found in Table 1.

3. Chamberlain's material

In the year of 1927 an investigation of the inheritance of left-handedness was performed by Chamberlain (1928). The material included 2177 families collected at the Ohio State University. The criterion of left-handedness was the writing hand. Besides the families collected at random, there are some families of type I in his report with which he came into contact through the newspapers. These families were added by Chamberlain to the material and used in his analysis, but thanks to his detailed report it is possible to exclude them from the calculations, which is necessary for a correct population-genetic analysis. Lacking the means to perform such an analysis, Chamberlain could only draw the conclusion, that handedness must be

TABLE 1. A COMPARISON BETWEEN VALUES CORRESPONDING TO DIFFERENT R-HYPOTHESES AND ACTUAL FREQUENCIES IN RIFE'S MATERIAL

Hypotheses	Type I		Type II		Type III		X^{2*}
	Left	Right	Left	Right	Left	Right	
(Actual va.)	6	5	34	140	151	1842	
r = 0.51	3.71	7.29	28.33	145.67	156.71	1836.29	3.714
r = 0.50	3.86	7.14	28.83	145.17	155.99	1837.01	3.112
r = 0.48	4.19	6.81	29.89	144.11	154.47	1838.53	2.030
r = 0.46	4.56	6.44	31.02	142.98	152.83	1840.17	1.149
r = 0.44	4.98	6.02	32.24	141.76	151.05	1841.95	0.490
r = 0.41	5.74	5.26	34.26	139.74	148.08	1844.92	0.089
r = 0.38	6.68	4.32	36.53	137.47	144.67	1848.33	0.697
r = 0.36	7.44	3.56	38.22	135.78	142.11	1850.89	2.057
r = 0.35	7.88	3.12	39.12	134.88	140.73	1852.27	3.252

* The number of degrees of freedom is one, as the information derived from the empirical data consists of five parameters.

hereditary. The frequencies found did not seem to correspond to any known Mendelian formula.

The P-generation contained 4354 individuals, 155 of which used their left hand in writing, and therefore recorded as left-handers (a = 0.0356).

The F-generation contained 7714 individuals, 367 of which were recorded as left-handers (b = 0.0476).

Family-type I (both parents left-handed) contained 25 children ($N_1 = 25$), 7 of which were recorded as left-handers ($X_1 = 7$). Equation (1) gives an r-value of 0.412.

Family-type II (one parent left-handed) contained 464 children ($N_2 = 464$), 53 of which were recorded as left-handers ($X_2 = 53$). Equation (2) gives an r-value of 0.393.

Family-type III (both parents right-handed) contained 7225 children ($N_3 = 7225$), 307 of which were recorded as left-handers ($X_3 = 307$). Equation (3) gives an r-value of 0.401.

The three r-determinations correspond as well in this material as in Rife's. The X^2-analyses show that r-values within the range of 0.49–0.33 are acceptable on the 5 per cent level, but values outside these limits make the hypothesis unacceptable. The smallest X^2-sum is obtained, when r equals 0.402, which may be looked upon as an estimation of R in the population from which Chamberlain's material was collected.

4. Ramaley's material

The oldest study dealt with in this context was performed by Francis Ramaley (1913). The material was collected at the University of Colorado, covering 305 families. There is no information in the report about the criterion used. On the basis of his figures Ramaley concluded that left-handedness is inherited as a Mendelian recessive, the frequencies of the genotypes assumed to be of the proportions 9:12:4 (the most probable constant proportion that he found among several enumerated,

containing, small, whole numbers). The reason behind this choice, giving 16 per cent recessive homozygotes, was the fact that he had determined the frequency of left-handedness in the F-generation to be 15.66 per cent. The but half as large percentage of left-handedness in the P-generation was explained to be caused by the parents having concealed their left-handedness to a great extent. The occurrence of a right-handed child in family-type I was thought to contradict the Mendelian hypothesis and could not be explained by Ramaley.

The P-generation contained 610 individuals, 49 of which were recorded as left-handers ($a = 0.0803$).

The F-generation contained 1130 individuals, 177 of which were recorded as left-handers ($b = 0.1566$).

Family-type I (both parents left-handed) contained 7 children ($N_1 = 7$), 6 of which were recorded as left-handers ($X_1 = 6$). Equation (1) gives an r-value of 0.427.

Family-type II (one parent left-handed) contained 170 children ($N_2 = 170$), 55 of which were recorded as left-handers ($X_2 = 55$). Equation (2) gives an r-value of 0.427.

Family-type III (both parents right-handed) contained 953 children ($N_3 = 953$), 116 of which were recorded as left-handers ($X_3 = 116$). Equation (3) gives an r-value of 0.425.

The three r-determinations give practically the same values. The X^2-analyses show that r-values within the range of 0.51–0.40 are acceptable on the 5 per cent level, but values outside these limits make the hypothesis unacceptable. The smallest X^2-sum is obtained, when r equals 0.427, which may be looked upon as an estimation of R in the population from which Ramaley's material was collected.

5. Discussion

In all these three population-genetic analyses of the inheritance of handedness the proposed model describes the empirical data completely. It thus seems probable that a dominant-recessive pair of alleles regulates the manifestation of handedness, the dominant gene conditioning right-handed behavior while left-handed behavior is found in the absence of this gene. According to the model the environment and eventual other genes cause a reduction of the manifestation of the recessive allele e.g. some recessive homozygotes express a behavior pattern which cannot be distinguished from the behavior of those individuals whose manual habits are determined by the dominant gene. Another possibility consistent with the model is given by the assumption that the right-handed behavior is determined by the total genotype, while left-handedness is caused by a diallelic gene, one allel of which has no influence of the total genotype, while the other as a double has a modifying effect which is counteracted by the environment and in consequence of the incomplete registering-methods not always observable. The well-known differences in strength of the left-handed tendencies may even in this case be explained by the influence exerted by other genes.[1]

[1] Dahlberg (1926) and others have found that identical twins frequently are discordant when it comes to handedness. Since identical twins must be assumed to have the same constitution, the dis-

TABLE 2. THE INHERITANCE OF HANDEDNESS IN THREE AMERICAN INVESTIGATIONS

	Ramaley 1913	Chamberlain 1928	Rife 1940
Parental Generation			
Total	610	4354	1374
Left-handers	49	155	72
a	0.0803	0.0356	0.0524
Filial Generation			
Total	1130	7714	2178
Left-handers	177	367	191
b	0.1566	0.0476	0.0877
Type I (left x left)			
N_1	7	25	11
X_1	6	7	6
r_1	0.427	0.412	0.401
Type II (left x right)			
N_2	170	464	174
X_2	55	53	34
r_2	0.427	0.393	0.414
Type III (right x right)			
N_3	953	7225	1993
X_3	116	307	151
r_3	0.425	0.401	0.439
Acceptance-limits	0.40–0.51	0.33–0.49	0.35–0.51
X^2-sum minimum	0.001	0.050	0.089
Estimation of R	0.427	0.402	0.410

The frequency of the R-gene for left-handedness lies between 40 and 43 per cent in all three studies. The mean of the established values is 41.3 per cent. The percentages of the various genotypes in the population of the United States corresponding to this value are DD: 34.5 per cent, DR: 48.5 per cent and RR: 17 per cent.

cordance has been taken to prove that hand preference could not be hereditary, at least not determined by the usual gene action. Dahlberg has assumed that genotypical asymmetries which are fairly common and to which he for the present wants to refer handedness, arise independently of the usual actions of the genes. In line with this opinion the discordance of the twins is an expression of such a mechanism. The twins are then considered each to be one half of the original, undivided embryo, divided into two asymmetrical parts by the twinning mechanism. The concordance among identical twins may be assumed to be due to the early timing of the parting of the embryo. That a mechanism of the type mentioned can be effective in twinning cannot be rejected. There are, however, no experimental or empirical findings indicating that the same or a similar mechanism could be the cause of the side-dominance in single-born children. For the present, it seems most correct to accept the possibility that handedness of identical twins can be changed by the twinning-mechanism. Deviations from expected Mendelian values in the population caused by such a mechanism are, however, most likely too small to be noticed in a population genetic analysis.

APPLICATION TO SCHIZOPHRENIA

The hereditary background of schizophrenia has been dealt with by Strömgren (1938), among others. He presented a hypothesis concerning its inheritance in his dissertation, which corresponds almost exactly to the theory proposed for left-handedness. Strömgren is of the opinion that predisposition for schizophrenia depends on the presence of a recessive gene but that the disorder appears only in a fraction of those who are homozygous for this gene. The incomplete manifestation is partly the effect of the environment and partly the effect of subordinate hereditary factors determining the degree of disposition for the disorder or the resistance to it. Strömgren's material is composed of disordered individuals and their families, thus giving no information of the frequency of schizophrenics within the main population to which these families belong. In order to be able to test the hypothesis, Strömgren is forced to make some additional assumptions regarding the probability of a recessive homozygote to become ill. This probability is designated m (Manifestationswahrscheinlichkeit) while the proportion of individuals becoming ill within a specific group is designated KE (Krankheitserwartung). With the support of these designations and the definition $m + n = 1$, Strömgren indicates how the frequency of ill individuals in the F-generation is regulated for different types of mating.

For an average population he gives the following formula:

$$KE = m\,r^2 \tag{4}$$

When one of the parents is ill:

$$KE = m\,\frac{dr + nr^2}{d^2 + 2dr + nr^2} \tag{5}$$

When none of the parents is ill:

$$KE = m\,\frac{d^2 + 2ndr + n^2r^2}{4d^2 + 4ndr + n^2r^2} \tag{6}$$

Since KE is a function of the two unknown values m and r in the empirical material (n and d may be expressed in these values) these formulae can not be used for an experimental test of the hypothesis of inheritance. Strömgren resorts therefore to using hypothetical values of m ("schätzungsweise veranschlagen"). Through inserting an arbitrary value of m into the equation for an average population, the KE of which he estimates to be 0.0075, he obtains a value of r (also arbitrarily). This r-value and other suitable m-values are then inserted into the equations (5) and (6), whereby he obtains an approximate correspondence to the frequency of illness in his own material.

This correspondence is not necessarily caused by the genetic constitution of his material since we do not know whether the inserted values of m and r really are valid in the material in question.

If we turn from the proband method to a population genetic method this is clarified further.[2] If we choose to use families selected at random from the main popula-

[2] The connections between Strömgren's equations and the general formulae are the following: In Strömgren's version (1) refers to the entire population. If we put $m = a/r^2$, KE is found to equal a

tion instead of ill individuals' families, m may be expressed directly in the relation between values obtainable from the material and r^2. The number of unknowns in the geno-statistical equations are thereby reduced to one, and the hypothesis may be tested according to the description.

A statistically valid test of Strömgren's hypothesis thus demands a material collected at random. If the expected acknowledgment is obtained the existence of a dominant recessive factor-pair may be established. The most important step for the establishment of the inheritance is thereby taken, although the differentiation of the recessive homozygotes remains to be clarified. Strömgren writes: "Mit einem Nachweis, dass ein rezessives Genpaar in der Schizofreniegenese obligat ist, ist die Frage des Erbganges ja nicht erschöpft; aber die übrigen in Betracht kommenden Gene werden doch von sekundärer Bedeutung sein, da sie für die Manifestation der Krankheit fördernd (oder hemmend) sein können, für sie aber nicht obligat (bzw, unvereinbar mit ihr) sind."

A HYPOTHESIS FOR DYSLEXIA

The most outstanding technical characteristic of a psychological genetic based on population-statistics is the consideration taken to varying manifestations of the genotypes (caused by environmental and measurement-technical influences). In order to test a hypothesis regarding the inheritance of a certain trait, we need formulae that include the influence of these factors. The hypothesis dealt with hitherto is clearly not the only one for which such formulae can be constructed.

In his study of dyslexia Hallgren (1950) has shown that the inheritance of the typical reading and writing difficulties most likely is autosomal, dominant monofactorial in character. Hallgren's investigation is based on proband methods which

---

in the P-generation. If we put $m = b/r^2$, KE is found to equal b in the F-generation. In the population-genetic theory these are the definitions of a and b. In Strömgren's version (2) is identical to the general formula for family-type II in the population-genetic theory. Since $KE = X_2/N_2$, and $m = a/r^2$ in the P-generation and b/r^2 in the F-generation, Strömgren's equation may be written

$$X_2/N_2 = \frac{b}{r^2} \cdot \frac{dr + (1 - a/r^2)r^2}{d^2 + 2dr + (1 - a/r^2)r^2} = \frac{b}{r^2} \cdot \frac{r - a}{1 - a}$$

In Strömgren's version (3) is not identical to the formula for family-type III in the population-genetic theory. It is true that it refers to the case when both parents are normal, but it presumes that the parents among their descendants has an ill individual, excluded from the material (the proband). This limits the genotypes of the parents to heterozygotes and normal recessive homozygotes, while the general formula also includes parents who are dominant homozygotes. The latter's contribution to the descendants of this family-type (d^2) will be recognized when the equation is evolved by substituting m as done previously. Formula (3) may thus be written

$$KE = \frac{b}{r^2} \cdot \frac{d^2 + 2(1 - a/r^2)dr + (1 - a/r^2)^2 r^2}{4d^2 + 4(1 - a/r^2)dr + (1 - a/r^2)^2 r^2} = \frac{b}{r^2} \frac{(d + (1 - a/r^2)r)^2}{(2d + (1 - a/r^2)r)^2}$$

$$= \frac{b}{r^2} \cdot \frac{(d + r - a/r)^2}{(2d + r - a/r)^2} = \frac{b}{r^2} \cdot \frac{(1 - a/r)^2}{(2d + r - a/r)^2} = \frac{b}{r^2} \cdot \frac{(r - a)^2}{(2dr + r^2 - a)^2}$$

$$= \frac{b}{r^2} \cdot \frac{(r - a)^2}{(1 - a - d^2)^2}$$

do not give complete information of the genetic constitution of the population. A doubt may be raised concerning Hallgren's way of assuming equality between the number of diagnosed cases of dyslexia and the number of corresponding genotypes, because there are good reasons to believe that some cases of dyslexia escape recording due to imperfect recording methods and because of environmental influences (the school and its education) working against the manifestation of the gene.

Hallgren's hypothesis may thus be formulated in a population-genetic way opening it to an experimental testing of the same type as the hypothesis of left-handedness and schizophrenia. This hypothesis may primarily be formulated in the following manner.

Dyslexia is inherited as an autosomal, monofactorial dominant. It is always manifested when the dominant gene is present in a double set. We must take into account that some of the heterozygotes will be recorded as normals in experimental work because of the environmental pressure working toward normal behavior, and because of incomplete recording. Heterozygotes will therefore be found both among those recorded as word-blind and those recorded as normals. Recessive homozygotes on the other hand, display only normal behavior. The analysis is carried out with the concepts used in the preceeding discussions. The proportion of trait-carriers in the P-generation is designated a, and in the F-generation, b. The proportions of the various genotypes in the P-generation will, according to the definitions, amount to

Recorded phenotype	Genotype	Proportion
Trait-carriers (dyslexic)	DD	d^2
	DR_d	$a - d^2$
Normals	DR_r	$1 - a - r^2$
	RR	r^2

The descendants in the different family-types are determined by the following combinations:

$$\begin{aligned}
&\text{Type I (trait-carrier)(trait-carrier)} = (DD+DR_d)(DD+DR_d)\\
&\text{Type II (trait-carrier)(normal)} = 2(DD+DR_d)(DR_r+RR)\\
&\text{Type III (normal)(normal)} = (DR_r+RR)(DR_r+RR)
\end{aligned}$$

The proportion of DD and DR in the three sub-groups of the F-generation are the following (for the fasciliation of writing the formulae, c is used as a substitute of $1 - a$):

Family-type I

The proportion of DD is

$$\frac{d^4 + \frac{1}{2}(a - d^2)d^2 + \frac{1}{2}(a - d^2)d^2 + \frac{1}{4}(a - d^2)^2}{a^2} = \frac{(a + d^2)^2}{4a^2}$$

The proportion of DR is

$$\frac{\frac{1}{2}(a - d^2)d^2 + \frac{1}{2}(a - d^2)d^2 + \frac{1}{2}(a - d^2)^2}{a^2} = \frac{a^2 - d^4}{2a^2}$$

Family-type II

The proportion of DD is

$$\frac{(c - r^2)d^2 + \frac{1}{2}(a - d^2)(c - r^2)}{2ac} = \frac{(c - r^2)(a + d^2)}{4ac}$$

The proportion of DR is

$$\frac{(c - r^2)d^2 + 2d^2r^2 + (a - d^2)(c - r^2) + (a - d^2)r^2}{2ac} = \frac{ac + d^2r^2}{2ac}$$

Family-type III

The proportion of DD is

$$\frac{\frac{1}{4}(c - r^2)}{c^2} = \frac{(c - r^2)^2}{4c^2}$$

The proportion of DR is

$$\frac{\frac{1}{2}(c - r^2)r^2 + \frac{1}{2}(c - r^2)r^2 + \frac{1}{2}(c - r^2)^2}{c^2} = \frac{c^2 - r^4}{2c^2}$$

With the knowledge of the proportions of the genotypes which give rise to trait-carriers in the F-generation, it is possible to formulate generalized expressions for the number of recorded trait-carriers in the three sub-groups of the F-generation. N_1, N_2, and N_3, designate the number of individuals in these groups, and X_1, X_2, and X_3, the number of recorded trait-carriers. All dominant homozygotes are recorded as trait-carriers, according to the hypothesis. The coefficient $\frac{b - d^2}{2dr}$ is employed to indicate the fraction of heterozygotes which are recorded as trait-carriers.

The three equations are:

Family-type I:

$$\left(\frac{(a + d^2)^2}{4a^2} + \frac{b - d^2}{2dr} \cdot \frac{a^2 - d^4}{2a^2}\right) \cdot N_1 = X_1 \tag{7}$$

Family-type II:

$$\left(\frac{(c - r^2)(a + d^2)}{4c^2} + \frac{b - d^2}{2dr} \cdot \frac{ac + d^2r^2}{2ac}\right) \cdot N_2 = X_2 \tag{8}$$

Family-type III:

$$\left(\frac{(c - r^2)^2}{4ac^2} + \frac{b - d^2}{2dr} \cdot \frac{c^2 - r^4}{2c^2}\right) \cdot N_3 = X_3 \tag{9}$$

Only d-values consistent with the original hypothesis may be accepted when putting these formulae to practical usage. When all heterozygotes are recorded as trait-carriers, a equals $d^2 + 2dr$, and when all heterozygotes are recorded as normals, a

equals d^2. On the basis of these expressions the theoretical limits for d (and r) can be determined to equal:

$$d_{max} = \sqrt{a},\ d_{min} = 1 - \sqrt{c}$$

When analysing an empirical material with the population genetic method already the determinations of the proportions of trait-carriers in the P- and F-generations give the limits within which the proportions of the two alleles may vary if the hypothesis is not to be rejected.

A collection of a material by which this modification of Hallgren's hypothesis could be tested, should not meet with any serious obstacles. It is not unlikely, however, that some other modification of the hypothesis, eventually may prove to be more correct. For instance, it is not impossible that the dominant homozygotes and heterozygotes are both receptive to the influences which reduce the mainfestation of the gene. The measuring instrument would then be likely to fail in recording both homo- and heterozygotes. If this is the case we must take into account that the genotype DD is represented also among those recorded as normals. The formulae will in that case be still more complicated.

REFERENCES

Chamberlain. H. D. 1928. The inheritance of left-handedness. *J. Hered.* 19: 557–559.

Dahlberg, G. 1926. Twin births and twins from a hereditary point of view. Stockholm: Bonniers.

Dahlberg, G. 1947. Mathematical methods for populations genetics. New York: S. Karger.

Hallgren, B. 1950. Specific Dyslexia. Köpenhamn: Einar Munksgaard.

Ramaley, F. 1913. Inheritance of left-handedness. *Amer. Nat.* 47: 730–738.

Rife, D. C. 1940. Handedness with special references to twins. *Genetics.* 25: 178–186.

Snyder, L. H. 1951. Old and New Pathways in Human Genetics. Amer. Jour. of Hum. Genet. 3: 1–16.

Stern, C. 1950. Principles of Human Genetics. San Francisco: Freeman & Co.

Strömgren, E. 1938. Beiträge zur psychiatrischen Erblehre. Köpenhamn: Einar Munksgaard.

Trankell, A. 1954. Ngoåt om möjligheterna till en psykologisk genetik. *Nordisk Psykologi.* 6: 1–19.

Trankell, A. 1955. Ytterligare något om psykologisk genetik. *Nordisk Psykologi.* 7: 65–72.

Sequential Tests for the Detection of Linkage[1]

NEWTON E. MORTON

University of Wisconsin

INFORMATION ON LINKAGE in man is accumulated as a succession of samples, each of which is typically small relative to the amount of data required to detect even moderately close linkage. The best method of analysis for such sequential samples, in the sense of requiring the least number of observations consistent with a given risk of error, has been found to be a sequential probability ratio test (Wald, 1947). It will now be shown that this test, in addition to minimizing the number of observations, is in other respects a useful method for the detection of linkage in man.

1. THE ASSUMPTIONS

Consider two gene loci, G and T, not necessarily on the same chromosome. An individual of genotype GG′ TT′ may be of either of two possible phases, GT/G′T′ or G′T/GT′, corresponding to his formation by the union of GT and G′T′ gametes, or of G′T and GT′ gametes. If the G and T loci happen to be on the same chromosome, these two phases correspond to the usual meanings of coupling and repulsion. In any case, the frequencies of the four types of gametes produced by this individual, if he is GT/G′T′, will be

$$\tfrac{1}{2}(1-\theta)\mathrm{GT},\quad \tfrac{1}{2}(1-\theta)\mathrm{G'T'},\quad \tfrac{1}{2}\theta\mathrm{GT'},\quad \tfrac{1}{2}\theta\mathrm{G'T},$$

whereas, if he is G′T/GT′, they will be

$$\tfrac{1}{2}\theta\mathrm{GT},\quad \tfrac{1}{2}\theta\mathrm{G'T'},\quad \tfrac{1}{2}(1-\theta)\mathrm{GT'},\quad \tfrac{1}{2}(1-\theta)\mathrm{G'T},$$

where θ is the probability of recombination between the two loci ($0 \leq \theta \leq 1$; nearly always, $\theta \leq 1/2$).

Now, a sufficient set of assumptions for a "linkage" test is the following:

1. The parental genotypes are known with certainty, except for phase.
2. The segregation ratios are not disturbed by incomplete penetrance or differential viability.
3. The method of ascertainment and selection of families is properly allowed for.

With this postulational basis, the null hypothesis to be tested is that "the three assumptions are correct and the recombination fraction in the population equals 1/2". Some of the alternative hypotheses are:

1. Incomplete penetrance or differential viability.
2. Biased ascertainment or selection of families.
3. Nonrandom segregation of nonhomologous chromosomes.
4. Co-existence of the two loci on the same chromosome (linkage).

Received May 28, 1955.

[1] Paper number 576 from the Department of Genetics.

Although a distinction between nonrandom chromosome segregation and linkage (which is presumably much the commoner of the two phenomena) will not be possible until the human linkage groups are better known, it should not be difficult to recognize the other disturbing factors in data that have been carefully collected and reported.

The above assumptions are rather stringent and must be examined in detail. Cases to be treated in this paper include incomplete ascertainment, uncertain parental genotypes, and incomplete penetrance.

No attempt will be made to treat "linkage" tests in which the basis of either character is not a single Mendelian factor. If the basis of one or both conditions is multifactorial or unknown, "linkage" is at best ambiguous and generally cannot be distinguished from any other phenotypic correlation which varies among families. The exploration of these complicated situations may be of some interest, but to include such characters on fancied "linkage" maps, as some authors have done, is to depreciate the linkage maps that have been determined with some precision in other organisms.

Since even the most conservative set of assumptions confounds linkage with other phenomena, the burden of proof is on the investigator who asserts that a particular example of linkage-like effects is evidence of true linkage. When two genes satisfy regular Mendelian ratios, however, it is convenient to denote such effects as linkage, with the assurance that this designation is rather precise, and that its precision will increase as the human linkage map is developed.

2. CURRENT TEST PROCEDURES

The three methods most commonly used to detect human linkage are the method of efficient scores (u scores), the Penrose sib-pair method, and the probability ratio method of Haldane and Smith (1947). Smith (1953) has recently shown that they are all really different forms of the nonsequential probability ratio test.

Valid scoring procedures were first applied to human linkage by Bernstein (1931), who showed that each family can be assigned a score whose sum, expected value, and variance provide a test of the null hypothesis in any body of data that is sufficiently large for the distribution of the total score to be nearly normal. Bernstein's scores were further developed by Hogben (1934) and Haldane (1934), but the evolution by Fisher of a maximum likelihood scoring procedure made these methods obsolete. Fisher (1935) was able to show that his u scores are more efficient than Bernstein's scores for all linkage intensities and are, in fact, fully efficient in the limit for loose linkage. Finney (1940 et seq.) has treated a great variety of cases by u scores, which are now commonly considered to be the method of choice whenever the amount of data is large and the families are not grouped into large pedigrees. However, u scores have certain disadvantages, some of which Smith (1953) has summarized as follows:

1. Although u scores are very easy to use when the parental genotype is completely known (except for phase), the calculation of the variance may be intractable when the parental genotypes are unknown. In large samples this can be circumvented by the use of a simple approximation (Smith, 1953).

2. The u scores are fully efficient only in the limit for loose linkage, which it is not practicable to detect. An ideal test would be efficient for moderate rather than loose linkage.

3. Information about linkage can be greatly increased by using data involving 3 or more generations. It is not feasible to extract this information by u scores.

4. The assumption of normality for the total score may be far from true for moderate sample sizes. Haldane (1946) has developed a normalizing transformation for such cases, and shown that in one instance an exact test fails to confirm the significance of a u score test.

The sib-pair method of Penrose has sometimes been recommended as an alternative to u scores when the parental genotypes are unknown. The investigations of Finney (1942) do not support this recommendation, since in his data the sib-pair method extracted only a small fraction of the information that could be obtained by u scores. However, when one of the test characters is a rare recessive trait, the sib-pair method fares somewhat better (Penrose, 1953). A serious disadvantage of the method is that it may be quite inexact when, as the current procedure requires, a family of size $s > 2$ is partitioned into all $s(s-1)/2$ possible pairs (Penrose, 1953; Smith, 1953). Smith (1953) has shown how a large-sample correction for non-independence of sib pairs may be applied, but its use destroys the principal advantage of the method, that of arithmetical simplicity. Finney (1941a) has pointed out that the Penrose sib-pair method is particularly sensitive to heterogeneity in gene frequencies when different populations are pooled. The sib-pair method can be applied to traits whose mode of inheritance is unknown, but then the term "linkage" is scarcely appropriate.

The probability ratio test of Haldane and Smith (1947) was devised to extract information from families and pedigrees without making the assumption of normality that is required by the maximum likelihood method. Their test depends on the theorem that the expected value of a probability ratio is 1 on the null hypothesis, regardless of the alternative hypothesis (Wald, 1947). Since this is true for any simple hypothesis, it must be true for any composite hypothesis, which is merely a weighted average of simple hypotheses such that the sum of the weights is 1. Let Λ be a probability ratio for the test of the null hypothesis that $\theta = 1/2$ against some alternative hypothesis. Then, on the null hypothesis, the inequality

$$\Lambda > A, \qquad (A > 1)$$

cannot occur with probability greater than $1/A$, since if it did, this in itself would be enough to raise the mean value $E(\Lambda)$ to 1, and therefore the occurrence of a value of Λ greater than A is at least as strong evidence against the null hypothesis as a significance level of $1/A$. Clearly this method of analysis has several advantages, among them its reliability in small as well as large samples, its dependence solely on elementary laws of probability, and the ease with which all kinds of families and pedigrees may be combined. However, the method is conservative, and a recent modification (average backward odds) is less efficient (Smith, 1953).

The three common methods of linkage detection in man do not exhaust the procedures that have been proposed, but of the current tests, the u statistics of Fisher

and Finney and the probability ratio method of Haldane and Smith are the best alternatives to sequential tests.

3. SEQUENTIAL TEST PROCEDURES

Let $f(y; \theta)$ denote the distribution of a random variable y, where θ is the recombination fraction and successive observations on y are indicated by $y_1, y_2, \cdots$, etc. The observation $y = 1$ signifies that $f(y; \theta)$ is of the form $f(1; \theta)$, and so on. For example, double backcross families of size 2 have two possible forms of the function $f(y; \theta)$, which may arbitrarily be specified by $y = 1$ and $y = 2$. Under the conditions of Section 8 below,

$$f(1; \theta) = \theta^2 + (1 - \theta)^2$$

$$f(2; \theta) = 2\theta(1 - \theta).$$

Thus, a particular sample of 3 independent sib pairs might be $y_1, y_2, y_3 = 2, 1, 2$, and the probability of this sample is $f(2; \theta)f(1; \theta)f(2; \theta)$.

Let H_0 be the null hypothesis that $\theta = 1/2$ and H_1 be the alternative hypothesis that $\theta = \theta_1$. The probability that a sample $y_1, y_2, \cdots, y_m$ is obtained is given by

$$p_{1m} = f(y_1 ; \theta_1) \cdots f(y_m ; \theta_1)$$

when H_1 is true, and by

$$p_{0m} = f(y_1 ; 1/2) \cdots f(y_m ; 1/2)$$

when H_0 is true. The sequential test (Wald, 1947) employs the probability ratio p_{1m}/p_{0m} and two positive numbers A and B, with $A > 1$ and $B < 1$. For purposes of practical computation it is much more convenient to work with the logarithm of this ratio rather than the ratio itself, since

$$\log \frac{p_{1m}}{p_{0m}} = \log \frac{f(y_1; \theta_1)}{f(y_1; 1/2)} + \cdots + \log \frac{f(y_m ; \theta_1)}{f(y_m ; 1/2)}.$$

Let z_i denote the i^{th} term in this sum, viz.,

$$z_i = \log \frac{f(y_i ; \theta_1)}{f(y_i ; 1/2)}.$$

The test procedure is carried out as follows, the quantities z_i ($i = 1, 2, \cdots$) being used: with each accession of data (consisting of one or more families or pedigrees), the cumulative sum $z_1 + \cdots + z_m$ is computed. If

$$\log B < z_1 + \cdots + z_m < \log A$$

the evidence on linkage is not decisive, and judgment with the preassigned significance level and power must be suspended until more data can be collected. If

$$z_1 + \cdots + z_m \geq \log A$$

there is significant evidence for linkage under the assumptions of the test. If

$$z_1 + \cdots + z_m \leq \log B$$

the recombination fraction is significantly greater than θ_1.

More data can always be used following a sequential test, either to estimate a significant linkage or to detect or exclude linkage in the range $\theta_1 < \theta \leq 1/2$, but this latter enterprise may be unprofitable if a stringent choice was made for θ_1.

The constants A and B are related to α, the probability of rejecting H_0 when H_0 is true (a Type I error), and β, the probability of rejecting H_1 when H_1 is true (a Type II error). In practice, two simple approximations are used to determine A and B:

$$A \cong \frac{1-\beta}{\alpha}$$

$$B \cong \frac{\beta}{1-\alpha}$$

Wald (1947) has shown that these approximations cannot result in any appreciable increase in the value of either α or β, and that they may be used to obtain expressions for the power function $P(\theta)$ and the average sample number function $E(n)$ of a sequential test. These two functions determine the best sequential test for a particular purpose and the extent of its superiority over nonsequential procedures. Requirements to impose on these functions are suggested by the probability distribution of θ.

4. THE PROBABILITY DISTRIBUTION OF THE RECOMBINATION FRACTION θ

Haldane and Smith (1947) have suggested "chiefly from a comparison with the known linkage values of *Drosophila*" that it may not be a bad approximation to assume that the recombination fraction for linked genes has a uniform distribution from 0 to 1/2. The distribution may also be arrived at more pedantically.

Consider a chromosome with genetic map length of L morgans, along which gene loci are distributed uniformly. We need not assume that the genes are distributed uniformly along the physical chromosome, only that their locations on the linkage map are so distributed. Choose two loci at random with locations C_1 and C_2, where C_1 is the first locus chosen. The quantity $w = |C_1 - C_2|$ is called the *map distance* between the two loci ($0 < w < L$). The cumulative density function of w may be represented on $(C_1/L, C_2/L)$ coordinates by the area within a unit square between the lines $w = C_2 - C_1$ and $w = C_1 - C_2$, or

$$F(w) = \frac{2}{L^2}\{\tfrac{1}{2}L^2 - \tfrac{1}{2}(L-w)^2\} = \frac{2Lw - w^2}{L^2}.$$

Kosambi (1944) has shown that the map distance w is related to the recombination fraction θ as

$$w = \tfrac{1}{4}\log\frac{1+2\theta}{1-2\theta}, \qquad 0 < \theta < \tfrac{1}{2}$$

assuming that the coincidence is 2θ. By this approximation

$$F(\theta) = \frac{\log \frac{1+2\theta}{1-2\theta}}{2L} - \frac{\left\{\log \frac{1+2\theta}{1-2\theta}\right\}^2}{16L^2}$$

and the probability distribution of θ for linked genes, gotten by differentiating $F(\theta)$, is

$$f(\theta) = \frac{2L - \frac{1}{2}\log \frac{1+2\theta}{1-2\theta}}{L^2(1-4\theta^2)}, \qquad 0 < \theta < \theta' < 1/2$$

$$= 0 \text{ elsewhere.}$$

The critical point θ' beyond which $f(\theta) = 0$ is determined by the equation

$$L = \tfrac{1}{4}\log \frac{1+2\theta'}{1-2\theta'} = \tfrac{1}{2}\tanh^{-1} 2\theta'$$

$$\therefore \theta' = \tfrac{1}{2}\frac{1-e^{-4L}}{1+e^{-4L}}.$$

We may verify that $f(\theta)$ is a density function over the interval 0 to θ';

$$F(\theta') = \frac{4L}{2L} - \frac{16L^2}{16L^2} = 1$$

since

$$\log \frac{1+2\theta'}{1-2\theta'} = 4L.$$

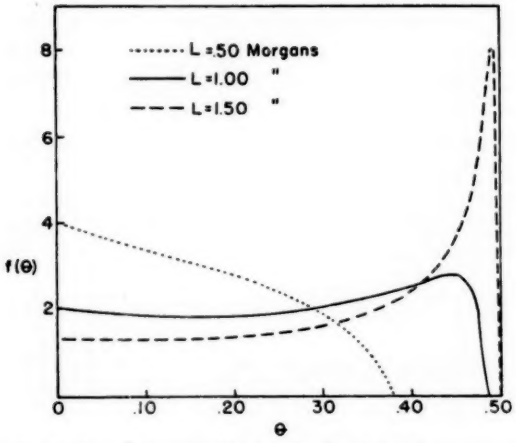

FIG. 1. The distribution of the recombination fraction θ for chromosomes of length L

TABLE 1.—THE DISTRIBUTION OF GENETIC MAP LENGTHS (L) IN DIFFERENT ORGANISMS

Source	L = .25	L = .50	L = .75	L = 1.00	L = 1.25	L = 1.50	L = 2.00	L = 2.50	L = 3.00	$\overline{L^2}/n(\overline{L})^2$
Drosophila[1]	—	—	1	2	—	—	—	—	—	.345
Corn (Zea)[2]	1	1	3	2	2	1	—	—	—	.117
Mouse[3]	—	15	—	48	—	46	13	3	2	.058

[1] Linkage map, neglecting the dot-like IVth chromosome, $L_{IV} = .002$ (Bridges and Brehme, 1944).

[2] Linkage map (Rhoades, 1950).

[3] Based on chiasma frequency in random chromosomes, assuming $L = \frac{\text{chiasma frequency}}{2}$ (Crew and Koller, 1932).

Recent data (Carter, 1955) suggest that the average value of L in the mouse is nearer to unity than here indicated, hence the distribution $g(\theta)$ in Figure 2 should presumably be even closer to uniformity.

Figure 1 shows $f(\theta)$ corresponding to different values of L. For chromosomes of length near unity (100 centimorgans) the distribution of θ is almost uniform. In fact, the recombination fraction has an exactly uniform distribution for chromosomes of unit genetic length according to the simple mapping function $\theta = w - \frac{1}{2}w^2$ $(0 < w < 1)$, for since $F(w) = 2w - w^2$, the distribution of θ is

$$F(\theta) = 2\theta, \qquad 0 < \theta < 1/2$$

$$f(\theta) = 2.$$

Actually chromosomes of unit length are nearly modal in the few higher organisms whose genetic maps are known. Table 1 gives the distribution of L for Drosophila, corn, and (very approximately) for the mouse. On the assumption of a uniform density of loci on the chromosome map, the probability distribution of the recombination fraction between two randomly chosen loci is

$$g(\theta) = \frac{\sum_L L^2 f(\theta)}{\sum_L L^2}.$$

Figure 2 shows that in all three species $g(\theta)$ is closely approximated by a uniform distribution, and that the greatest departure from this approximation is for values of θ close to 1/2, which in practice could seldom be distinguished from independent assortment. The distribution $g(\theta)$ is probably much the same in man, where the average genetic length, based on mean chiasma frequency, may be close to unity (Schultz, unpublished; cited by Neel, 1949).

Table 1 may also be used to compute the probability ϕ that two randomly chosen loci be on the same chromosome. If the number of loci per chromosome is proportional to L,

$$\phi = \frac{\sum L^2}{(\sum L)^2} = \frac{\overline{L^2}}{n(\overline{L})^2}$$

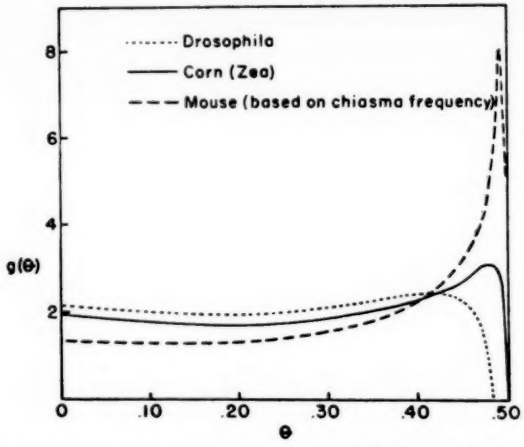

FIG. 2. The distribution of the recombination fraction θ for linked genes in three different species

where n is the haploid number of chromosomes. If all chromosomes are of equal length, $\phi = 1/n$, and for the organisms tabulated this turns out to be a good approximation. In Drosophila, neglecting the dot-like IVth chromosome, n = 3, $\phi = .345$; in corn, n = 10, $\phi = .117$; in the mouse, n = 20, $\phi = .058$, or $\phi = .064$ if pachytene length is proportional to L (Slizynski, 1949). In man, with 23 autosomes, the frequency of autosomal linkage may reasonably be taken as $\phi = .05$, so that the distribution of recombination values in man may be approximated as follows:

$$g(\theta) = 2\phi = .10 \qquad 0 < \theta < 1/2$$

$$= 1 - \phi = .95 \qquad \theta = 1/2$$

$$= 0 \qquad \text{elsewhere.}$$

5. THE CHOICE OF A SEQUENTIAL TEST

The validity of a sequential test does not depend on the accuracy of these approximations, but they do suggest criteria by which a suitable sequential test may be selected. We are especially anxious to avoid the assertion that two genes are linked when in fact they are not, since a misleading linkage map is worse than no linkage map at all. One source of linkage-like effects can be nearly eliminated by considering only pairs of loci which satisfy our assumption that the expected segregation ratios for both loci are realized in the population sampled. However, cases of apparent linkage will still be made up in part of true linkages, in part of Type I errors. If the prior probability of linkage is $\phi = .05$, then the posterior probability that a case of apparent linkage be a Type I error is

$$\rho = \frac{\alpha(1 - \phi)}{\alpha(1 - \phi) + \phi\overline{P}} = \frac{19\alpha}{19\alpha + \overline{P}}.$$

where $\bar{P}$ is the average power of the test, or the probability of detecting linkage when it is present. R. S. Krooth (personal communication) has termed ρ the *reliability* and $\bar{P}$ the *sensitivity* of a linkage test. Calculations of ρ for different values of α and $\bar{P}$ show that the usual values of α are inadequate in this problem, and that for the posterior probability of a Type I error to be less than .05, α must be about .002 when $\bar{P} = .95$, .001 when $\bar{P} = .60$ and .0005 when $\bar{P} = .20$ (cp. Haldane, 1934).

Having placed the requirement on α that it be small enough to reduce the posterior probability of a Type I error to .05, we impose a second condition on the power function of the test. To be at all useful, the test must have a power close to unity for values of θ near zero. We are at liberty to choose θ_1, the formal alternative to $\theta_0 = 1/2$, as near to 1/2 as we please, and the only adverse effect of this choice is to increase the average sample number. On this reasoning it seems appropriate to let θ_1 take the largest value which is likely to give a significant result in a practicably large body of data, and to consider the average sample number function a basis for the selection of a sequential test.

As an application of this argument, consider four sequential test procedures defined by the relations

(1) $\theta_1 = .05, \quad A = 2000, \quad B = .01, \quad \theta_0 = 1/2$

(2) $\theta_1 = .10, \quad A = 1000, \quad B = .01, \quad \theta_0 = 1/2$

(3) $\theta_1 = .20, \quad A = 1000, \quad B = .01, \quad \theta_0 = 1/2$

(4) $\theta_1 = .30, \quad A = 1000, \quad B = .01, \quad \theta_0 = 1/2$

and assume that the data consist entirely of double backcross sibships of size 2, sampled under the conditions of §8 below. The probability can take only the value $f(1; \theta) = \theta^2 + (1 - \theta)^2$, corresponding to a sib pair that is either concordant in both traits or discordant in both, and $f(2; \theta) = 2\theta(1 - \theta)$, which corresponds to a sib pair that is concordant in one trait and discordant in the other. Following Wald (1947) and assuming that the excess over the boundaries at the termination of the test can be neglected, we obtain a good approximation to the power function $P(\theta)$ by solving two equations for various values of h

$$P(\theta) = \frac{1 - B^h}{A^h - B^h}$$

and

$$\sum_y f(y; \theta) \left[\frac{f(y; \theta_1)}{f(y; 1/2)}\right]^h = 1.$$

From the power function, again neglecting the excess over the boundaries, we obtain the average sample number function as

$$E_\theta(n) = \frac{P(\theta) \log A + [1 - P(\theta)] \log B}{E_\theta(z)}$$

where

$$E_\theta(z) = \sum_y f(y; \theta) \log \left[\frac{f(y; \theta_1)}{f(y; 1/2)}\right].$$

In particular,

$$E_{\theta_1}(n) = \frac{(1-\beta)\log A + \beta \log B}{E_{\theta_1}(z)}$$

and

$$E_{\theta_0}(n) = \frac{\alpha \log A + (1-\alpha)\log B}{E_{\theta_0}(z)} \qquad \text{(Wald, 1947).}$$

The power functions and average sample number functions for the four test procedures are plotted in figures 3 and 4, the information from which is summarized in table 2. All four tests have power greater than .99 for values of θ less than .05 and power less than .03 for values of θ greater than .40. In the intervening range, the first test has good power at $\theta = .10$, the second is moderately good at $\theta = .20$, the third has appreciable power at $\theta = .30$, and the fourth is good for all values of θ less than $\theta = .35$. The value of α has been taken so as to keep the posterior probability of a Type I error (ρ) nearly constant and less than .05, provided that the assumptions of the previous sections are satisfied. The average power $\overline{P}$ increases from .28 to .71, and the average sample number, which represents the cost of this gain in power, increases from 10 to 355.

The investigator will probably seldom have need for sequential tests outside the above range. A test so insensitive as not to detect virtually all cases of close linkage ($\theta < .05$) is of little use, while an increase in sensitivity much beyond $\theta_1 = .30$ requires a prohibitively large average sample number: for example, when $\theta = 1/2$, the test $\theta_1 = .40$, A = 1000, B = .01 requires an average sample number of 5700 double backcross sib pairs.

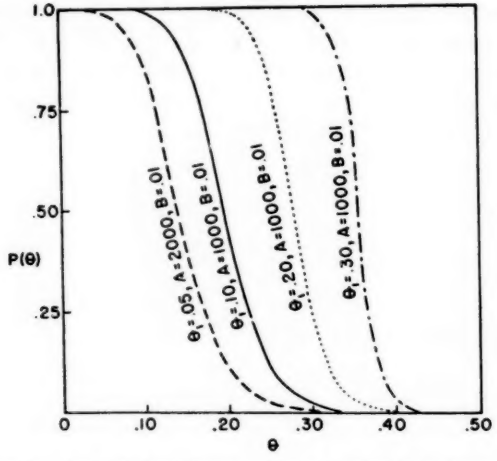

FIG. 3. The power function $P(\theta)$ for different values of θ_1. Double backcross sibships of size 2

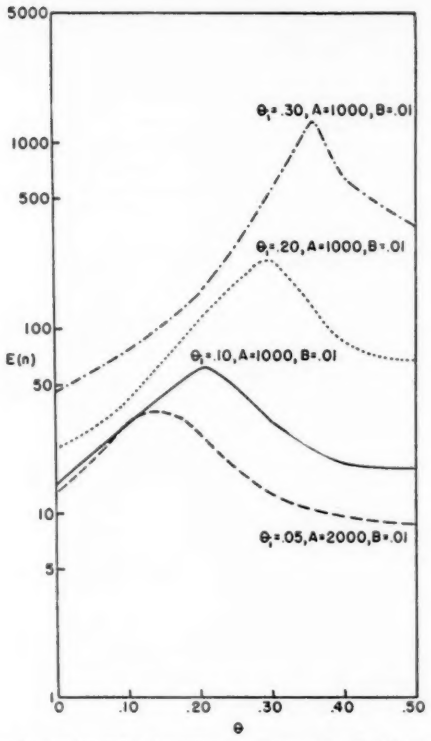

FIG. 4. The average sample number E(n) for different values of θ_1. Double backcross sibships of size 2

6. THE NUMBERS OF OBSERVATIONS REQUIRED BY FIXED-SAMPLE-SIZE TESTS AND SEQUENTIAL TESTS

The exposition so far has considered criteria by which a sequential test may be chosen, and has suggested a battery of four tests which should be adequate for most purposes. We still require, however, to select among these procedures and, more immediately, to determine whether a sequential test is so superior to current fixed-sample-size tests in efficiency, computational simplicity, or exactness that the choice of a sequential test has more than academic interest.

For a start, we may calculate the number of independent double backcross sib pairs required by current tests of strength (α, β). In the case of u statistics there are two possible scores, 1 and -1, with frequencies $\theta^2 + (1 - \theta)^2$ and $2\theta(1 - \theta)$ respectively (Finney, 1940). The expected value of the score is $\mu_\theta = (1 - 2\theta)^2$, with variance $\sigma_\theta^2 = (1 - \mu_\theta)(1 + \mu_\theta)$. (Note that these symbols designate the expected value and variance of the score, not of θ.) If the sample size is small, it may be estimated by trial and error from a table of the cumulative binomial distribution, using the parameters $p_1 = 2\theta_1(1 - \theta_1)$ and $p_0 = 2\theta_0(1 - \theta_0) = 1/2$. If the sample

TABLE 2.—CHARACTERISTICS OF FOUR SEQUENTIAL TESTS

θ_1 = the formal alternative to the null hypothesis that $\theta = \frac{1}{2}$.
α = the probability of rejecting the null hypothesis when $\theta = \frac{1}{2}$.
β = the probability of accepting the null hypothesis when $\theta = \theta_1$.
$P(\theta)$ = the probability of detecting linkage when the true recombination fraction is θ.
$\overline{P}$ = the probability of detecting linkage when θ is uniformly distributed between 0 and $\frac{1}{2}$.
ρ = the probability that a significant "linkage" be a Type I error.
$\overline{E(n)}$ = the average number of double backcross sibships of size 2 required to terminate the test.

θ_1	α	β	$P(\theta)$				$\overline{P}$	ρ	$\overline{E(n)}$
			$\theta = .10$	$\theta = .20$	$\theta = .30$	$\theta = .35$			
.05	.0005	.01	.86	.10	.006	.002	.28	.032	10
.10	.001	.01	.99	.46	.02	.006	.39	.046	19
.20	.001	.01	> .999	.99	.23	.025	.56	.032	68
.30	.001	.01	> .999	> .999	.99	.64	.71	.026	355

$$\overline{P} = 2\int_0^{1/2} P(\theta)\, d\theta$$

$$\rho \simeq \frac{19\alpha}{19\alpha + \overline{P}}$$

$$\overline{E(n)} \simeq .10\int_0^{1/2} E_\theta(n)\, d\theta + .95E_{1/2}(n)$$

is sufficiently large, the distribution of the sample mean will be nearly normal, and the following conditions will determine $n(\alpha, \beta)$, the required sample number:

$$G\left[\frac{d - \mu_{\theta_0}}{\sigma_{\theta_0}/\sqrt{n}}\right] = 1 - \alpha$$

$$G\left[\frac{d - \mu_{\theta_1}}{\sigma_{\theta_1}/\sqrt{n}}\right] = \beta$$

where d is a preassigned constant defining the critical region of the test and

$$G(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}\, dx.$$

If we let t_0 be the value for which $G(t_0) = 1 - \alpha$, and t_1 be the value for which $G(t_1) = \beta$, and observe that $\mu_{\theta_0} = 0$ and $\sigma_{\theta_0} = 1$, then the two conditions may be written as

$$\sqrt{n}\, d = t_0$$

$$\sqrt{n}(d - \mu_{\theta_1}) = t_1\sqrt{(1 - \mu_{\theta_1})(1 + \mu_{\theta_1})}.$$

Solving the above equations, we obtain

$$n = n(\alpha, \beta) = \left[\frac{t_0 - t_1\sqrt{(1 - \mu_{\theta_1})(1 + \mu_{\theta_1})}}{\mu_{\theta_1}}\right]^2.$$

If this expression is not an integer, then, as in all formulae determining fixed sample size, $n(\alpha, \beta)$ is the smallest integer in excess (Wald, 1947).

In the case of the probability ratio test of Haldane and Smith (1947), there are two possible values of the logarithm of the probability ratio, namely

$$z' = \log\left[\frac{f(1; \theta_1)}{f(1; 1/2)}\right] = \log(2 - 4\theta_1 + 4\theta_1^2)$$

and

$$z'' = \log\left[\frac{f(2; \theta_1)}{f(2; 1/2)}\right] = \log[4\theta_1(1 - \theta_1)].$$

The expected value of z_θ is $\mu_\theta = z' - 2\theta(1 - \theta)(z' - z'')$, with variance

$$\sigma_\theta^2 = (z')^2 - 2\theta(1 - \theta)(z' - z'')(z' + z'') - (\mu_\theta)^2.$$

The first condition determining the sample size is

$$\sum z = \log(1/\alpha),$$

and if n is sufficiently large, the second condition becomes

$$\frac{\sum z - n\mu_{\theta_1}}{\sqrt{n}\,\sigma_{\theta_1}} = t_1.$$

Solving for n, we obtain

$$n = n^*(\hat{\alpha}, \beta) = \left[\frac{\sqrt{t_1^2\sigma_{\theta_1}^2 + 4\mu_{\theta_1}\log(1/\alpha)} - t_1\sigma_{\theta_1}}{2\mu_{\theta_1}}\right]^2.$$

For the Haldane-Smith test the true significance level $\hat{\alpha}$ is less by a varying amount than the nominal level α, so that in this respect the test is conservative. Smith (1953) calculated that the median $\hat{\alpha}$ is approximately $\alpha/10$ for $\hat{\alpha} = .001$. The error of the normal approximation in determining $n(\alpha, \beta)$ and $n^*(\hat{\alpha}, \beta)$ is in the opposite direction, since the alternative distribution is skewed toward $\theta_0 = 1/2$, and therefore β and n tend to be underestimated. This error is negligible unless n is very small, and in table 3, which gives the results of these calculations, the smallest value of $n(\alpha, \beta)$ is in close agreement with an exact determination from the cumulative binomial distribution.

TABLE 3.—THE AVERAGE SAMPLE NUMBER E(n) FOR A SEQUENTIAL TEST, COMPARED WITH THE FIXED SAMPLE NUMBERS REQUIRED BY THE FISHER-FINNEY U SCORE TEST, $n(\alpha, \beta)$, AND THE HALDANE-SMITH PROBABILITY RATIO TEST, $n^*(\hat{\alpha}, \beta)$

n = the required number of double backcross sibships of size 2.

θ_1	α	β	E(n)		$n(\alpha, \beta)$	$n^*(\hat{\alpha}, \beta)$
			$\theta = \frac{1}{2}$	$\theta = \theta_1$		
.05	.0005	.01	9	20	34	49
.10	.001	.01	18	31	59	89
.20	.001	.01	67	103	214	328
.30	.001	.01	355	529	1,134	1,740
.40	.001	.01	5,700	8,546	18,324	28,420

The conclusions from table 3 are quite simple and consistent. Of the fixed-sample-size tests, u statistics require only about 2/3 as many observations for a given risk of error as the Haldane-Smith probability ratio test. If, in view of the conservatism of the latter test, a value of α ten times as large is used, the number of observations required by the test is intermediate between $n(\alpha, \beta)$ and $n^*(\hat{\alpha}, \beta)$, and is still appreciably in excess of the sample size required by the u score test.

Although the superiority of the u score test over the Haldane-Smith probability ratio test is marked, the superiority of the sequential test is even more striking. When the alternative hypothesis is true, the sequential test requires only about 1/2 as many observations as a u score test of the same strength, and when the null hypothesis is true (as it usually will be), the sequential test requires less than 1/3 as many observations as the u score test. Similar savings in the number of observations have been found for other distributions by Wald (1947) and Bross (1952).

For the detection of linkage we have knowledge that the user of a sequential test does not ordinarily have, in that the approximate parameter distribution is known, and we may calculate a mean sequential sample number $\overline{E(n)}$ averaged over this distribution (table 2). Over the range of tests considered, the mean sample number required by a sequential test of strength (α, β) is less than 1/3 the number required by a u score test of the same strength.

7. CLASSIFICATION OF FACTORS, MATINGS, AND METHODS OF SAMPLING

In view of the considerable saving in observations indicated in the last section, sequential tests would seem to be the method of choice for the detection of linkage. For practical use, the determination of probabilities must be extended to families of different types and sizes. We first require a few definitions.

Consider two loci, G and T, which are to be tested for linkage. The genetic characters which are determined by these loci may be divided into four classes. These are:

1. Recessive abnormalities, such as albinism. The symbols G,g or T,t will be used for factors of this class.

2. Common recessives, such as the gene for the inability to taste phenylthiocarbamide. Symbols G,g or T,t will also be used here.

3. Factors without dominance, the heterozygote being distinguishable from both homozygotes. Sicklemia and the MN blood groups are examples of this class. The letters G_1, G_2 or T_1, T_2 will be used for such factors.

4. "Dominant" abnormalities, such as ovalocytosis. The normal homozygote is exceedingly rare (in most cases never having been observed), and all abnormal persons are therefore assumed to be heterozygous. The symbol G_1 or T_1 will be used for the normal allele, G_2 or T_2 for the abnormal factor.

For a family to give information on linkage, neither parent may be GG or TT and at least one parent must be doubly heterozygous. An informative mating is termed a double backcross, a single backcross, or a double intercross according to whether the other parent is doubly homozygous, singly heterozygous, or doubly heterozygous. Since the phase of linkage is unknown, the probability for a double or single backcross will consist of two terms, one for each possible phase of the

doubly heterozygous parent, and the probability for a double intercross will consist of three terms, corresponding to the possibilities that both parents are in coupling, both in repulsion, or that one is in coupling and the other in repulsion. We shall assume that the two phases are at equilibrium in the population, a condition that should nearly always be closely approximated, except perhaps after recent hybridization. On the null hypothesis this assumption is of course supererogatory.

It rarely happens that families selected for a linkage study are effectively a random sample from the general population. Usually families are selected first on the basis of the character determined by the "main" locus and are tested afterwards for the character determined by the "test" locus. There are three methods of selecting families on the basis of the main character (Bailey, 1951):

1. Selection through the parents or grandparents, without consideration of the children. The sampling of families is effectively random, and in families of a given mating type and size, the distribution of the number of children manifesting the main character is a complete binomial series (*complete* selection).

2. Selection through the children themselves, with complete selection of affected individuals. In families of a given mating type and size, the distribution of the number of children manifesting the main character is a truncated binomial series, with the first term missing (*truncate* selection).

3. Selection through the children, with incomplete selection of affected individuals. The distribution of affected individuals in sibships of a given mating type and size is not a truncated binomial, since families with large numbers of affected children are more likely to be ascertained than families with a smaller number of abnormals (*arbitrary* selection). This is the usual method of selection for recessive abnormalities and a not uncommon method of selection for "dominant" abnormalities and rare factors without dominance.

Except in cases of gross ascertainment bias, the test character is never subject to incomplete selection of affected individuals (method 3).

It should be noted that these three methods of selecting families for analysis subsume the rejection of some classes of ascertained families. The fundamental attribute of each type of selection is the distribution to which it gives rise, regardless of how the families were detected. For example, with recessive genes the propositus is sometimes an affected parent mated to a normal dominant, who may be either homozygous or heterozygous. A mating of a dominant parent is called "certain" if there is at least one recessive child (in which case the dominant parent must be heterozygous), and is called "doubtful" otherwise. Sampling is by method 1 or 2, according to whether doubtful families are included or rejected. The method of ascertainment is the same in both cases, but the method of selection is different, and determines the proper method of analysis.

8. BOTH CHARACTERS SELECTED THROUGH THE PARENTS (COMPLETE SELECTION). PARENTAL GENOTYPES KNOWN, BOTH PARENTS TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

Unless there is no dominance for either character, some of the families will usually be of uncertain parental genotype. If these doubtful families are analysed separately

TABLE 4.—MATINGS SCORED WITH z_1. DOUBLE BACKCROSSES AND SINGLE BACKCROSSES WITH NO DOMINANCE IN THE INTERCROSS FACTOR

$s = a + b + c + d$

Parental genotype	Mating Type	Progeny Phenotype				Uninformative Progeny
		a	b	c	d	
Gg Tt × gg tt	1	G T	G t	g T	g t	—
Gg T_1T_2 × gg T_1T_1	2	G T_1	G T_1T_2	g T_1	g T_1T_2	—
G_1G_2 Tt × G_1G_1 tt	3	G_1 T	G_1 t	G_1G_2 T	G_1G_2 t	—
Gg T_1T_2 × gg T_1T_2	4	G T_1	G T_2	g T_1	g T_2	T_1T_2
G_1G_2 Tt × G_1G_2 tt	5	G_1 T	G_1 t	G_2 T	G_2 t	G_1G_2
G_1G_2 T_1T_2 × G_1G_1 T_1T_1	6	G_1 T_1	G_1 T_1T_2	G_1G_2 T_1	G_1G_2 T_1T_2	—
G_1G_2 T_1T_2 × G_1G_1 T_1T_2	7	G_1 T_1	G_1 T_2	G_1G_2 T_1	G_1G_2 T_2	T_1T_2
G_1G_2 T_1T_2 × G_1G_2 T_1T_1	8	G_1 T_1	G_1 T_1T_2	G_2 T_1	G_2 T_1T_2	G_1G_2

Frequency	a	b	c	d	Total
Coupling 1	$1-\theta$	θ	θ	$1-\theta$	2
Repulsion 1	θ	$1-\theta$	$1-\theta$	θ	2
Total	1	1	1	1	4

$$z_1 = \log \frac{f(y;\theta_1)}{f(y;\frac{1}{2})} = \log 2^{s-1}[\theta_1^{a+d}(1-\theta_1)^{b+c} + \theta_1^{b+c}(1-\theta_1)^{a+d}]$$

(see §12), then the methods of this section are appropriate to the certain families. If the doubtful families are rejected, the certain families should be analysed by the methods of §§9–10.

Neglecting multiple allelism, the possible kinds of certain families may be grouped into 5 classes, which by the method of u scores have 3 essentially different scores and 2 derived scores (Finney, 1940). In sequential tests the same classes exist. The scores in a sequential test are "lods", or logarithms of the probability ratio, the five functional forms of which may be denoted by z_1, z_2, z_3, z_4, and z_5, in exact correspondence with the u_{11}, u_{31}, u_{33}, $2u_{31}$, and $2u_{11}$ scoring types of Finney.

Tables 4–8 give the possible certain matings and the lod scores appropriate to them. Matings scored with z_1 (table 4) comprise double backcrosses and those single backcrosses in which there is no dominance for the intercross factor. There is thus a one-to-one correspondence between progeny genotype and phenotype for both loci. Note that some progeny have probabilities that are independent of the recombination fraction and phase, and therefore give no information on linkage. Matings scored with z_2 (table 5) are single backcrosses with dominance in the intercross factor. Matings scored with z_3 (table 6) are double intercrosses with dominance in both factors. Most matings of common occurrence are scored with the z_1, z_2, or z_3 lods, of which the z_1 type is much the most informative.

The two remaining scoring types are of particular interest because the u score method omits progeny from which information is extracted by the lod scores. Matings scored with z_4 (table 7) are double intercrosses with dominance in only one factor. There are six progeny phenotypes, the last two of which have probabilities that are

TABLE 5.—MATINGS SCORED WITH z_2. SINGLE BACKCROSSES WITH DOMINANCE IN THE INTERCROSS FACTOR

Parental genotype	Mating Type	Progeny phenotype				Uninformative Progeny
		a	b	c	d	
Gg Tt × Gg tt	9	G T	g T	G t	g t	—
Gg Tt × gg Tt	10	G T	G t	g T	g t	—
Gg T_1T_2 × Gg T_1T_1	11	G T_1	g T_1	G T_1T_2	g T_1T_2	—
G_1G_2 Tt × G_1G_1 Tt	12	G_1 T	G_1 t	G_1G_2 T	G_1G_2 t	—

Frequency	a	b	c	d	Total
Coupling 1	$2-\theta$	θ	$1+\theta$	$1-\theta$	4
Repulsion 1	$1+\theta$	$1-\theta$	$2-\theta$	θ	4
Total	3	1	3	1	8

$$z_2 = \log \frac{f(y;\theta_1)}{f(y;\frac{1}{2})} = \log \frac{2^{a-1}}{3^{a+c}} [(2-\theta_1)^a \theta_1^b (1+\theta_1)^c (1-\theta_1)^d + (1+\theta_1)^a (1-\theta_1)^b (2-\theta_1)^c \theta_1^d]$$

TABLE 6.—MATINGS SCORED WITH z_3. DOUBLE INTERCROSSES WITH DOMINANCE IN BOTH FACTORS

Parental genotype	Mating Type	Progeny phenotype				Uninformative Progeny
		a	b	c	d	
Gg Tt × Gg Tt	13	G T	G t	g T	g t	—

Frequency	a	b	c	d	Total
G T/g t × G T/g t 1	$3-2\theta+\theta^2$	$\theta(2-\theta)$	$\theta(2-\theta)$	$(1-\theta)^2$	4
G T/g t × G t/g T 2	$2+\theta-\theta^2$	$1-\theta+\theta^2$	$1-\theta+\theta^2$	$\theta(1-\theta)$	8
G t/g T × G t/g T 1	$2+\theta^2$	$1-\theta^2$	$1-\theta^2$	θ^2	4
Total	9	3	3	1	16

$$z_3 = \log \frac{f(y;\theta_1)}{f(y;\frac{1}{2})} = \log \frac{4^{a-1}}{9^a 3^{b+c}} \Big[(3-2\theta_1+\theta_1^2)^a \theta_1^{b+c} (2-\theta_1)^{b+c} (1-\theta_1)^{2d} + 2 (2+\theta_1-\theta_1^2)^a \cdot (1-\theta_1+\theta_1^2)^{b+c} \theta_1^d (1-\theta_1)^d + (2+\theta_1^2)^a (1-\theta_1^2)^{b+c} \theta_1^{2d} \Big]$$

linear functions of $\theta(1-\theta)$, whereas the other four types include terms which are not linear in $\theta(1-\theta)$, like θ^2. When $\theta \to 1/2$, the deviation of $\theta(1-\theta)$ from 1/4 is vanishingly small compared with the deviation of θ^2 from 1/4, and the last two classes contribute almost no information on linkage. It is not surprising, therefore, that when the probability is expanded in powers of $1-2\theta$, and the cubic and higher terms neglected, the appropriate u score is a function of only the first four classes (Finney, 1940). Since loose linkage ($\theta \to 1/2$) is never in practice distinguished from non-linkage ($\theta = 1/2$), the important consideration is that the information contributed by the neglected progeny (which constitute 1/2 of the total children) is not negligible when θ is small.

TABLE 7. MATINGS SCORED WITH z_4. DOUBLE INTERCROSSES WITH DOMINANCE IN ONE FACTOR

$s = a + b + c + d + e + f$

Parental genotype	Mating Type	Progeny phenotype						Uninformative Progeny
		a	b	c	d	e	f	
Gg T_1T_2 x Gg T_1T_2	14	G T_1	g T_1	G T_2	g T_2	G T_1T_2	g T_1T_2	—
G_1G_2 Tt x G_1G_2 Tt	15	G_1 T	G_1 t	G_2 T	G_2 t	G_1G_2 T	G_1G_2 t	—

Frequency		a	b	c	d	e	f	Total
Coupling x coupling	1	$1 - \theta^2$	θ^2	$\theta(2 - \theta)$	$(1 - \theta)^2$	$2(1 - \theta + \theta^2)$	$2\theta(1 - \theta)$	4
Coupling x repulsion	2	$1 - \theta + \theta^2$	$\theta(1 - \theta)$	$1 - \theta + \theta^2$	$\theta(1 - \theta)$	$1 + 2\theta - 2\theta^2$	$1 - 2\theta + 2\theta^2$	8
Repulsion x repulsion	1	$\theta(2 - \theta)$	$(1 - \theta)^2$	$1 - \theta^2$	θ^2	$2(1 - \theta + \theta^2)$	$2\theta(1 - \theta)$	4
Total		3	1	3	1	6	2	16

$$z_4 = \log \frac{f(y;\theta_1)}{f(y;\frac{1}{2})} = \log \frac{4^{s-1}}{3^{a+c+e}2^{e+f}} \{2^{c+f}\theta_1^{2b+c+f}(1-\theta_1)^{a+2d+f}(1+\theta_1)^a(2-\theta_1)^c(1-\theta_1+\theta_1^2)^e + 2\theta_1^{b+d}(1-\theta_1)^{b+d}(1-\theta_1+\theta_1^2)^{a+c}(1+2\theta_1-2\theta_1^2)^c(1-2\theta_1+2\theta_1^2)^f + 2^{c+f}\theta_1^{a+2d+f}(1-\theta_1)^{2b+c+f}(1+\theta_1)^c(2-\theta_1)^a(1-\theta_1+\theta_1^2)^e\}$$

TABLE 8. MATINGS SCORED WITH z_5. DOUBLE INTERCROSSES WITH NO DOMINANCE IN EITHER FACTOR

$s = a + b + c + d + e + f + g + h + i$

Parental genotype	Mating Type	Progeny phenotype									Uninformative Progeny
		a	b	c	d	e	f	g	h	i	
G_1G_2 T_1T_2 x G_1G_2 T_1T_2	16	G_1 T_1	G_1 T_2	G_2 T_1	G_2 T_2	G_1G_2 T_1	G_1G_2 T_2	G_1 T_1T_2	G_2 T_1T_2	G_1G_2 T_1T_2	—

Frequency		a	b	c	d	e	f	g	h	i	Total
Coupling x coupling	1	$(1 - \theta)^2$	θ^2	θ^2	$(1 - \theta)^2$	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2(1 - 2\theta + 2\theta^2)$	4
Coupling x repulsion	2	$\theta(1 - \theta)$	$\theta(1 - \theta)$	$\theta(1 - \theta)$	$\theta(1 - \theta)$	$1 - 2\theta + 2\theta^2$	$1 - 2\theta + 2\theta^2$	$1 - 2\theta + 2\theta^2$	$1 - 2\theta + 2\theta^2$	$4\theta(1 - \theta)$	8
Repulsion x repulsion	1	θ^2	$(1 - \theta)^2$	$(1 - \theta)^2$	θ^2	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2(1 - 2\theta + 2\theta^2)$	4
Total		1	1	1	1	2	2	2	2	4	16

$$z_4 = \log \frac{f(y;\theta_1)}{f(y;1/2)} = \log \frac{4^{s-1}}{2^{w+i}} \left\{2^w\theta_1^{2v+w}(1-\theta_1)^{2u+w}(1-2\theta_1+2\theta_1^2)^i + 2^{i+1}\theta_1^{u+v+i}(1-\theta_1)^{u+v+i}(1-2\theta_1+2\theta_1^2)^w + 2^w\theta_1^{2u+w}(1-\theta_1)^{2v+w}(1-2\theta_1+2\theta_1^2)^i\right\}$$

$u = a + d$
$v = b + c$
$w = e + f + g + h$

Matings scored with z_5 (table 8) are double intercrosses with no dominance in either factor. The lod score is based on 9 distinguishable progeny classes, the last 5 of which contribute no information when $\theta \to 1/2$, and are therefore neglected in computing the u scores (Finney, 1940). When θ is small, however, the information contained in these children (which constitute 3/4 of the progeny) is no longer negligible.

9. ONE CHARACTER SELECTED THROUGH THE PARENTS (COMPLETE SELECTION), THE OTHER THROUGH THE CHILDREN (INCOMPLETE SELECTION). PARENTAL GENOTYPES KNOWN, BOTH PARENTS TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

For convenience we may denote the factor that is selected through the children by G,g, G_1, or G_2, and the factor selected through the parents by T, t, T_1, or T_2. The method of this section is appropriate only if families of doubtful parental genotype with regard to the T locus are not rejected (section 12); the selection of the G factor is arbitrary.

In a family of size s let there be s_1 children of one G type, say G, and s_2 of the other ($s_1 + s_2 = s$). The prior probability of the family will be designated by $f(y;\theta)$ and the conditional probability by $f(y;\theta \mid s_1)$. Then

$$f(y;\theta \mid s_1) = \frac{f(y;\theta)}{P(s_1,s_2)}$$

where $P(s_1,s_2)$ is the probability measure of the selected class of families. Since the two characters are selected independently, and the probabilities which are pooled in $P(s_1,s_2)$ are complementary, $P(s_1,s_2)$ is independent of θ and of the phase of linkage and cancels when the probability ratio is formed. Thus the probability ratio and the lod score derived from it have the convenient property of being invariant with respect to biased sampling of one character only, and families selected in this way are scored just as if both characters had been ascertained through the parents (Smith, 1953).

10. BOTH CHARACTERS SELECTED THROUGH THE CHILDREN, COMPLETE SELECTION OF AFFECTED INDIVIDUALS (TRUNCATE SELECTION). PARENTAL GENOTYPES KNOWN, BOTH PARENTS TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

Families in which the parental genotype is unknown for either factor are rejected. The condition on both factors makes the marginal distribution of the selected families a function of θ, and the methods of the previous sections require modification. There are three types to be considered, corresponding to the z_1, z_2, and z_3 scoring types. We shall suppose that the selected factors are g and t, since only matings in which both characters are common recessives are likely to be selected in this way.

(1) The z_1 scoring type (Mating 1)

The distribution of the selected families is

$$f(y;\theta \mid g, t) = \frac{f(y;\theta)}{P(g,t)}$$

where P(g,t) is the probability that a mating of this type have at least one g and one t child. To satisfy this condition, it is sufficient that $c + d \neq 0$ and $b + d \neq 0$. Therefore,

$$P(g,t) = 1 - P(c + d = 0) - P(b + d = 0) + P(b + c + d = 0).$$

$$\text{But } P(c + d = 0) = \sum_{a=0}^{s} \binom{s}{a} \left\{ \frac{1}{2} \left(\frac{1-\theta}{2} \right)^a \left(\frac{\theta}{2} \right)^{s-a} + \frac{1}{2} \left(\frac{1-\theta}{2} \right)^{s-a} \left(\frac{\theta}{2} \right)^a \right\}$$

$$= (1/2)^s = P(b + d = 0)$$

$$\text{and } P(b + c + d = 0) = P(a = s) = \frac{1}{2} \left\{ \left(\frac{1-\theta}{2} \right)^s + \left(\frac{\theta}{2} \right)^s \right\}, \text{ and so}$$

$$P(g,t) = \frac{2^s - 2 + \frac{1}{2}\theta^s + \frac{1}{2}(1-\theta)^s}{2^s}.$$

It follows that

$$\log \frac{f(y;\theta_1 \mid g,t)}{f(y;1/2 \mid g,t)} = \log \frac{f(y;\theta_1)}{f(y;1/2)} + \log \frac{P(g,t;1/2)}{P(g,t;\theta_1)}$$

$$= z_1 + c_1$$

$$\text{where } c_1 = \log \frac{2^s - 2 + (1/2)^s}{2^s - 2 + \frac{1}{2}\theta_1^s + \frac{1}{2}(1-\theta_1)^s}.$$

Thus the lod score in this case, and in general, is simply the score appropriate to random sampling plus a correction factor which is determined by the method of selection. The factor c_1 is exactly analogous to $-\epsilon_s$ in the theory of u scores (Finney, 1940).

(2) The z_2 scoring type (Matings 9 and 10)

Using the same notation as before, we find that

$$\log \frac{f(y;\theta_1 \mid g,t)}{f(y;1/2 \mid g,t)} = z_2 + c_2$$

$$\text{where } c_2 = \log \frac{4^s - 2^s - 3^s + (3/2)^s}{4^s - 2^s - 3^s + \frac{1}{2}(2-\theta_1)^s + \frac{1}{2}(1+\theta_1)^s}.$$

(3) The z_3 scoring type (Mating 13)

$$\log \frac{f(y;\theta_1 \mid g,t)}{f(y;1/2 \mid g,t)} = z_3 + c_3.$$

$$c_3 = \log \frac{4^s - 2(3)^s + (9/4)^s}{4^s - 2(3)^s + \frac{1}{4}(3 - 2\theta_1 + \theta_1^2)^s + \frac{1}{2}(2 + \theta_1 - \theta_1^2)^s + \frac{1}{4}(2 + \theta_1^2)^s}.$$

11. BOTH CHARACTERS SELECTED THROUGH THE CHILDREN, ONE COMPLETELY (TRUNCATE SELECTION), THE OTHER INCOMPLETELY (ARBITRARY SELECTION). PARENTAL GENOTYPES KNOWN, BOTH PARENTS TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

Let the character with arbitrary selection be denoted by g or G_2, and let t denote the character with truncate selection. The family is ascertained through the

G factor and then tested for the T factor, with rejection of families in which there is not at least one t child. (If these families are not rejected, or if there is no dominance in the T factor, see §9.) Occasionally the method of incomplete ascertainment of the G factor may be known exactly, but the simplest and most reliable procedure is to consider the distribution of the families with the G factor fixed, so that the method of selection does not enter into the argument (Finney, 1940).

A. Dominance in the G factor (G,g type)

Let there be s_1 children of type G and s_2 of type g ($s_1 + s_2 = s$). The distribution of selected families is

$$f(y;\theta \mid s_1,s_2,t) = \frac{f(y;\theta)}{P(s_1,s_2,t)}$$

where $P(s_1,s_2,t)$ is the probability measure of selected families of this class. Note that $s_2 = 0$ implies ascertainment of the G factor through the parents or uninformative children, hence the s_1,s_2 method of scoring is not appropriate unless $s_2 > 0$ or the viability of the G,g types is abnormal.

(1A) The z_1 scoring type (Mating 1)

$$P(s_1,s_2,t) = P(s_1,s_2) - P(s_1,s_2\,,\, b + d = 0)$$

$$P(s_1,s_2) = k\binom{s}{s_1}(1/2)^{s_1}(1/2)^{s_2}$$

$$P(s_1,s_2,\, b + d = 0) = P(a = s_1\,,\, c = s_2) = k\binom{s}{s_1}\left\{\frac{1}{2}\left(\frac{\theta}{2}\right)^{s_1}\left(\frac{1-\theta}{2}\right)^{s_2} + \frac{1}{2}\left(\frac{\theta}{2}\right)^{s_2}\left(\frac{1-\theta}{2}\right)^{s_1}\right\}.$$

Therefore,

$$P(s_1,s_2,t) = k\binom{s}{s_1}(1/2)^s\{1 - \tfrac{1}{2}\,\theta^{s_1}(1-\theta)^{s_2} - \tfrac{1}{2}\,\theta^{s_2}(1-\theta)^{s_1}\},$$

where k is a selection factor dependent only on s_1 and s_2 and

$$\log\frac{f(y;\theta_1 \mid s_1,\, s_2,\, t)}{f(y;1/2 \mid s_1,s_2,t)} = z_1 + e_1$$

where

$$e_1 = \log\frac{1 - (1/2)^s}{1 - \tfrac{1}{2}\,\theta_1^{s_1}(1-\theta_1)^{s_2} - \tfrac{1}{2}\,\theta_1^{s_2}(1-\theta_1)^{s_1}}.$$

(2A) The z_2 scoring type (Mating 9)

$$\log\frac{f(y;\theta_1 \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_2 + e_2$$

$$e_2 = \log\frac{3^{s_1}[1 - (1/2)^s]}{3^{s_1} - \tfrac{1}{2}\,(2-\theta_1)^{s_1}\theta_1^{s_2} - \tfrac{1}{2}\,(1+\theta_1)^{s_1}(1-\theta_1)^{s_2}}.$$

(3A) The z_2 scoring type (Mating 10)

$$\log \frac{f(y;\theta_1 \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_2 + d_2$$

$$d_2 = \log \frac{2^s - (3/2)^s}{2^s - \frac{1}{2}(2-\theta_1)^{s_1}(1+\theta_1)^{s_2} - \frac{1}{2}(1+\theta_1)^{s_1}(2-\theta_1)^{s_2}}.$$

(4A) The z_3 scoring type (Mating 13)

$$\log \frac{f(y;\theta \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_3 + e_3$$

$$e_3 = \log \frac{3^{s_1}[1 - (3/4)^s]}{3^{s_1} - \frac{1}{4}(3 - 2\theta_1 + \theta_1^2)^{s_1}\theta_1^{s_2}(2-\theta_1)^{s_2} - \frac{1}{2}(2+\theta_1-\theta_1^2)^{s_1}(1-\theta_1+\theta_1^2)^{s_2} - \frac{1}{4}(2+\theta_1^2)^{s_1}(1-\theta_1^2)^{s_2}}.$$

B. Incomplete dominance in the G factor (G_1G_2 type)

Rare "dominants" and a few characters lacking dominance (sicklemia, thalassemia) are sometimes selected incompletely in this way. This situation was not considered by Finney (1940).

(1B) The z_1 scoring type (Mating 3)

Let s_1 be the number of G_1 children, and s_2 be the number of G_1G_2 children. Then the probability ratio is the same as for type 1A above, and

$$\log \frac{f(y;\theta_1 \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_1 + e_1.$$

(2B) The z_1 scoring type (Mating 5)

If the family is selected through a G_1G_2 child, then there is random sampling for the informative progeny, and the method of section 9 applies. If selection is through an informative G_1 or G_2 child, then

$$\log \frac{f(y;\theta_1 \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_1 + e_1,$$

where s_1 is the number of G_1 children and s_2 the number of G_2 children.

(3B) The z_2 scoring type (Mating 12)

Let there be s_1 children of type G_1 and s_2 children of type G_1G_2. The probability ratio is the same as for 3A above, and

$$\log \frac{f(y;\theta_1 \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_2 + d_2.$$

(4B) The z_4 scoring type (Mating 15)

Let there be s_1 children of type G_1, s_2 of type G_1G_2, and s_3 of type G_2 ($s_1 + s_2 + s_3 = s$). Then

$$\log \frac{f(y;\theta_1 \mid s_1,s_2,s_3,t)}{f(y;1/2 \mid s_1,s_2,s_3,t)} = z_4 + e_4$$

$$e_4 = \log \frac{1 - (3/4)^s}{1 - \frac{1}{4}(1 - \theta_1^2)^{s_1}(1 - \theta_1 + \theta_1^2)^{s_2}[\theta_1(2 - \theta_1)]^{s_3} - (1/2)^{s_2+1}(1 - \theta_1 + \theta_1^2)^{s_1+s_3} \cdot (1 + 2\theta_1 - 2\theta_1^2)^{s_2} - \frac{1}{4}[\theta_1(2 - \theta_1)]^{s_1}(1 - \theta_1 + \theta_1^2)^{s_2}(1 - \theta_1^2)^{s_3}}.$$

This completes the analysis of the matings in tables 4–8. These include all the scoring types of Finney (1940), who used 3 essentially different scores, 2 derived scores, 7 score corrections, and 12 essentially different information functions. For the same matings, the probability ratio method requires only 5 scores and 7 correction factors. The development of the probability ratio scores is extremely simple and may easily be extended to more complex cases, such as multiple allelism, uncertain parental genotypes, and pedigree data. To facilitate numerical analysis of the matings that have been treated so far, the scores for small families are given in tables 10–18.

12. PARENTS OF UNKNOWN GENOTYPE, BOTH PARENTS TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

Parental heterozygosity for recessive factors can be established by the observation of recessive children, in the absence of which a family without pedigree information is termed "doubtful". Information may still be extracted from these families, provided that the population gene frequencies are known and that mating is at random with respect to the doubtful locus. We have seen in §9 that when families are selected through the parents for the test factor, and doubtful families are not rejected, then no score correction is needed for families of known parental genotype regardless of how the main character is selected. Matings doubtful for the main character may also be analysed.

In connection with the doubtful families it will be convenient to introduce a few new symbols. Let p_t denote the frequency of the t gene and p_g the frequency of the g gene. Occasionally children will not be scorable for linkage, either because they are uninformative or because they are incompletely tested. If these children are tested for the doubtful character, they give information about the parental genotypes and should enter into the present calculations. Let S be the number of scored and unscored children which are tested for the doubtful character, in contradistinction to s, the number of children which are scored for linkage. As an example of the general procedure, we shall develop scores for the "doubtful" analogues of the z_1 scoring type.

(*1*) *Families doubtful for the t factor* (*Matings 1, 3, 5*)

All children are of type T. The prior probabilities for homozygosity and heterozygosity of the T parent are $(1 - p_t)^2$ and $2p_t(1 - p_t)$, and the conditional probabilities for the children are

$$(1/2)^s \text{ and } \tfrac{1}{2}\{\theta^a(1 - \theta)^c + \theta^c(1 - \theta)^a\}(1/2)^S$$

respectively. Therefore,

$$\log \frac{f(y;\theta_1)}{f(y;1/2)} = \log \frac{2^{S-s} - p_t\{2^{S-s} - \theta_1^a(1 - \theta_1)^c - \theta_1^c(1 - \theta_1)^a\}}{2^{S-s} - p_t\{2^{S-s} - (1/2)^{s-1}\}}.$$

(2) Families doubtful for the g factor (Matings 1, 2, 4)

All children of type G. The probability ratio is the same as for the previous type, except for the substitution of p_g for p_t and b for c.

$$\log \frac{f(y;\theta_1)}{f(y;1/2)} = \log \frac{2^{s-s} - p_g\{2^{s-s} - \theta_1^a(1-\theta_1)^b - \theta_1^b(1-\theta_1)^a\}}{2^{s-s} - p_g\{2^{s-s} - (1/2)^{s-1}\}}.$$

(3) Families doubtful for the g and t factors (Mating 1)

All children of type GT. The GT parent may be GGTT, GgTT, GGTt, or GgTt, only the last of which is informative. The lod score is

$$\log \frac{f(y;\theta_1)}{f(y;1/2)} = \log \frac{2^{s-1} - (2^{s-1}-1)(p_g+p_t) + p_g p_t\{2^{s-1} - 2 + \theta_1^s + (1-\theta_1)^s\}}{2^{s-1} - (2^{s-1}-1)(p_g+p_t) + p_g p_t\{2^{s-1} - 2 + (1/2)^{s-1}\}}.$$

The scoring system for the doubtful families may easily be extended to the analogues of the z_2, z_3, and z_4 scoring types. However, the application of these scores is quite tedious in the absence of ancillary tables for each of the common test factors and, more important, the doubtful families have in practice been found to contribute relatively little information on linkage. Finney found in one example that scoring doubtful families for the ABO locus increased the available amount of information by only 5%, and he advised that "for a preliminary investigation of a linkage, scoring may well be confined to the certain families" (Finney, 1940). This policy, besides reducing the labor in linkage detection, has the further advantage of making linkage tests independent of the mating system and the population gene frequencies. Unless the data are extremely valuable, it seems best to score only the certain families, using where necessary the correction factors of §§10–11.

13. One or both parents not directly tested. Complete penetrance, no natural selection

The extraction of information from untested parents by the method of u scores involves considerable algebraic manipulation and heavy arithmetic. Finney (1941b) has treated a few special cases and Smith (1953) has suggested an approximation for use in large samples. Fortunately the probability ratio method is so simple that *ad hoc* computation is always feasible, although the calculations are still tedious.

Suppose first that all ascertained families with untested parents are to be analysed, subject to the condition that families are sampled through the parents for both characters or that they are sampled through the parents for one character and the parental genotypes for the other character are known. On these assumptions the method of ascertainment does not affect the calculation, which consists in enumerating all parental genotypes which could give rise to F, the family in question, and then computing from the population gene frequencies and the assumption of random mating the prior probabilities of the mating types, say $P(M_1)$, $P(M_2)$, $\cdots$ etc. The conditional probabilities, $P(F \mid M_1)$, $P(F \mid M_2)$, $\cdots$ etc. are then calculated. Finally, the score for linkage is computed as

$$\log \frac{f(y;\theta_1)}{f(y;1/2)} = \log \frac{\sum_i P(M_i)P(F \mid M_i,\theta_1)}{\sum_i P(M_i)P(F \mid M_i,1/2)}$$

which of course is zero if none of the conditional probabilities is a function of θ.

These calculations are straightforward but time-consuming, and the investigator of human linkage would be well-advised to test both parents whenever possible. Full information cannot be recovered from incomplete records, although large families, whose scores are dominated by the conditional probabilities, are nearly as informative as if both parents had been tested. If instances of incomplete parental testing are not too common, no great amount of information will be lost by rejecting families with incomplete parental records. Alternatively, the scoring of incomplete records may be restricted to families whose parental genotypes can be inferred with certainty. In this case the linkage test is independent of gene frequencies and the mating structure of the population, considerable labor is saved, and at least some large families with only one tested parent will be included in the analysis. The score for the families whose parental genotypes are inferred is $z + C$, where z is the score appropriate to complete selection with both parents tested and C is a correction factor dependent on the method of sampling and inference. There are many special cases for C, all of which are easily treated *ad hoc* by the elementary methods used in §§10–11.

14. NATURAL SELECTION AND INCOMPLETE PENETRANCE

Genetic main factors with incomplete penetrance or low viability may still be used for linkage studies if we assume that the test factor is fully penetrant, viable, sampled at random through the parents or through complete selection of affected children, and that the viability and penetrance of the main factor are independent of the test factor.

For example, suppose the main factor is fully penetrant but so subvital that many affected progeny die before examination. On the above assumptions, it is still proper to test linkage by the methods of §§9 and 11, and the probabilities of Type I and Type II errors remain unaltered. Notice that no assumption need be made about the constancy of viability among families, either in the detection or estimation of linkage.

Again, suppose that the main gene is incompletely penetrant, with no assumptions made about viability or ascertainment. We shall assume that the main factor is so rare that all matings will be backcrosses if the main factor is a rare "dominant" or intercrosses if the main factor is a rare recessive. Given the above conditions on the test factor, the probability of a Type I error when the methods of §§9 and 11 are used will not be changed, regardless of whether penetrance is variable or not, but the power of the test will decrease very greatly when penetrance is low. In this case estimation of the penetrance will improve the power of the test, without affecting the probability of a Type I error.

In practice, the distinction between loose linkage to the main factor and linkage to viability or penetrance modifiers may be difficult to make, and therefore only tests of close linkage have much value when viability or penetrance is irregular. Even with such tests the rigorous justification of the assumption that the test factor does not influence the viability or penetrance of the main factor is extremely difficult, and may well be attempted only for tests which indicate a significant "linkage". Proof that the main and test factors are distributed independently in the general population, the absence of a correlation between the test phenotype of affected

parents and affected progeny, constant penetrance, and homogeneity of the linkage value give supporting evidence for the hypothesis of linkage, while contrary observations suggest alternative explanations. Knowledge of the exact method of ascertainment is helpful in detecting irregularities, especially with rare recessive factors. All these problems are particularly acute when the test factor is extremely complex, and great difficulties have been encountered in attempts to distinguish linkage when sex is used as the test factor (Harris, 1948; Mohr, 1954). Even with less fundamental test traits, a significant "linkage" effect requires special scrutiny when the penetrance or viability of the main factor is low. If the test factor also behaves irregularly, the difficulties in linkage detection are vastly increased.

15. THE COMBINATION OF DATA

In §§5–6 the properties of the sequential probability ratio test were illustrated on the simplifying assumption that the data consist entirely of double backcross sibships of size 2, and it was shown that for this case the sequential test is very much superior to alternative procedures. In practice, linkage data in man comprise a mixture of family sizes and mating types, the frequencies of which vary among pairs of loci and are usually unspecified. We shall now show that this ignorance does not affect the important properties of the sequential test.

Let $k = 1, 2, \cdots$, denote a particular mating type and family size, $f_k(y;\theta)$ be the conditional distribution for the k^{th} type of data, and p_k be the prior probability of this type of data. Consider only sampling procedures for which p_k and $f_k(y;\theta)$ are independent of the stage of sampling. Then clearly the distribution $p_k f_k(y;\theta)$ is of the stationary type treated by Wald and all the important results of his sequential theory apply. In particular, it has been shown that of all tests with the same risk of error (α, β), the sequential probability ratio test requires on the average fewest observations, and that the Type I and Type II risks are approximately

$$\alpha = \frac{1 - B}{A - B}$$

$$\beta = \frac{B(A - 1)}{A - B},$$

these approximations being very good when the excess of $\sum z$ over the boundary $\log A$ or $\log B$ is negligible. This condition is satisfied if $| E(z) |$ and the standard deviation σ_z of z are sufficiently small, as in practice they usually will be. In any case the optimum character of the sequential test holds exactly (Wald and Wolfowitz, 1948).

Although the existence of a stationary distribution $p_k f_k(y;\theta)$ is sufficient for the proof of the above remarks, it is not necessary that the p_k be known to carry out the test. For the p_k are independent of θ, and therefore the probability ratio

$$\prod \frac{p_k f_k(y;\theta_1)}{p_k f_k(y;\theta_0)}$$

is identical with the ratio

$$\prod \frac{f_k(y;\theta_1)}{f_k(y;\theta_0)}.$$

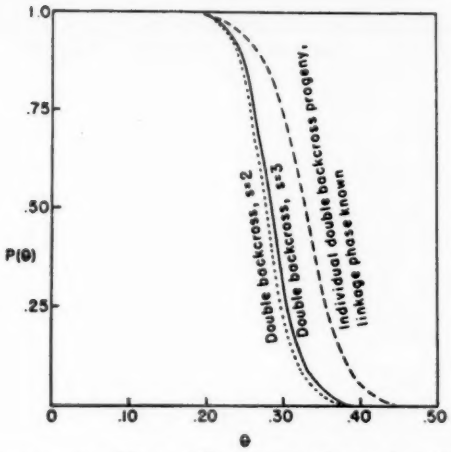

FIG. 5. The power function $P(\theta)$ for different types of data. A = 1000, B = .01, θ_1 = .20.

Determination of the p_k is necessary only if it is desired to find the power function and average sample number function of a sequential test, but this is of secondary importance so long as there is some basis for the choice of a particular test and we know that the sequential test on the average leads to a saving in the number of observations.

To choose a sequential test, it is convenient to have a rough notion of the average power of alternative tests. The power function depends on the distribution p_k, but the risks (α, β) do not, and this limits the possible fluctuation of the power function. Figure 5 shows a typical power function for three different types of data. The power function and the average power do not seem to be so highly variable as to jeopardize the control over Type I errors demanded for the idealized case in §5. In particular, it still seems appropriate to choose an unusually small value of α, of the order of .001.

The choice of θ_1 for a sequential test is largely determined by the average sample number on the null hypothesis, since (1) for randomly chosen loci the null hypothesis will usually be true and (2) the number of observations that can be tolerated is not narrowly bounded, so that random excesses over the expected number will usually not be a serious annoyance. A rough correspondence between expected sample number and amount of information may be established as follows.

Let n be the number of families required to terminate the test in mixed data and n_k be the number of families required for the test in data entirely of the k^{th} type. Let E(z) denote the expected value of z in mixed data and $E(z_k)$ the expected value of z in the k^{th} type of data. Also let c be a fixed value of k. Then on the null hypothesis

$$E(n)E(z) = \alpha \log A + (1 - \alpha) \log B = E(n_c)E(z_c)$$

and

$$E(n_c) = E\left\{\sum_{i=1}^{E(n)} \left[\frac{E(z_k)}{E(z_c)}\right]_i\right\},$$

TABLE 9.—THE EFFICIENCY OF DIFFERENT TYPES OF DATA IN DOUBLE BACKCROSS SIB-PAIR EQUIVALENTS

$\theta = \frac{1}{2}$

Scoring Type	θ_1 .05	θ_1 .20	θ_1 .40	u score information $\theta_1 \to \frac{1}{2}$
A. Families of size s, phase unknown				
z_1, s = 2	1.0	1.0	1.0	1.0
z_2, s = 2	.1	.1	.1	.2
z_3, s = 2	.1	.1	.1	.2
z_1, s = 5	3.8	5.3	8.8	10.0
z_3, s = 5	.5	.5	.5	.8
z_1, s = 10	9.7	14.7	33.1	45.0
B. Single progeny, phase known				
double backcross	1.6	3.2	25.4	—
single backcross	.5	1.0	8.5	—
double intercross, coupling, both factors dominant	1.0	1.8	12.3	—

where i = 1, 2, $\cdots$, E(n) denotes successive observations from the distribution $p_k f_k(y;\theta)$. If we let c designate double backcross sibships of size 2, then the ratio $E(z_k)/E(z_c)$ may be called the *double backcross sib-pair equivalent* on the null hypothesis. It has the property that if $E(n_c)$ is the average number of double backcross sib-pairs required by a certain test when $\theta = \theta_0 = 1/2$, then $E(n_c)E(z_c)/E(z_k)$ is the average number of families of type k required for the same test, assuming in both cases that the excess over the boundaries at the termination of the test can be neglected. Furthermore, for small families $E(z_k)/E(z_c)$ is of the same order as the information weight k in Finney's (1940) system of u scores (table 9). It follows that if S is the number of units of u score information that can be obtained with "reasonable" effort, then S is an estimate of $\sum E(z_k)/E(z_c)$ and $E(n_c)$ also, and this correspondence may serve as a rough guide in the selection of a sequential test. If S is about 10, θ_1 should be chosen to be .05, since $E(n_c) = 9$ for $\theta_1 = .05$. Similarly, if S is about 70, θ_1 should be taken as .20, if S is as much as 350, θ_1 may be .30, and only if S is about 6000 should θ_1 be .40. For linkage of two common test factors (ABO, Rh, MN), S may be as much as 6000, and for two less common test factors (Le, Lu, P, Fy blood groups), S may be 350. In most other cases S is probably smaller than 100, and θ_1 should be chosen accordingly. If it turns out that S has been considerably underestimated, a second test with a larger value of θ_1 will not increase α beyond tolerable limits.

The restriction of the sampling procedure to stationary distributions has proscribed a valid sampling method that in some respects seems desirable. All types of data might be collected at the beginning of sampling and whenever linkage is suggested, but when there is no suggestion of linkage it would seem economical to investigate only highly informative families for which the double backcross sib-pair equivalent is large. This makes p_k dependent on $\sum z$, but $f_k(y;\theta)$ is not affected and the probability is still one that the procedure will eventually terminate. It is of course essential that data be reported without regard for whether they indicate linkage or not. Wald (1947) has shown that the postulated kind of dependence does

TABLE 10

z_1

s	a + d	b + c	θ_1				
			.05	.10	.20	.30	.40
2	0	2	.2577	.2148	.1335	.0645	.0170
	1	1	−.7212	−.4437	−.1938	−.0757	−.0177
	2	0	.2577	.2148	.1335	.0645	.0170
3	0	3	.5353	.4654	.3181	.1703	.0492
	1	2	−.7212	−.4437	−.1938	−.0757	−.0177
	2	1	−.7212	−.4437	−.1938	−.0757	−.0177
	3	0	.5353	.4654	.3181	.1703	.0492
4	0	4	.8140	.7201	.5171	.2979	.0940
	1	3	−.4636	−.2289	−.0603	−.0113	−.0007
	2	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	3	1	−.4636	−.2289	−.0603	−.0113	−.0007
	4	0	.8140	.7201	.5171	.2979	.0940
5	0	5	1.0927	.9753	.7200	.4358	.1486
	1	4	−.1860	.0217	.1242	.0945	.0315
	2	3	−1.4425	−.8874	−.3876	−.1514	−.0355
	3	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	4	1	−.1860	.0217	.1242	.0945	.0315
	5	0	1.0927	.9753	.7200	.4358	.1486
6	0	6	1.3715	1.2306	.9238	.5784	.2106
	1	5	.0927	.2764	.3233	.2222	.0763
	2	4	−1.1848	−.6726	−.2541	−.0870	−.0184
	3	3	−2.1637	−1.3311	−.5815	−.2272	−.0532
	4	2	−1.1848	−.6726	−.2541	−.0870	−.0184
	5	1	.0927	.2764	.3233	.2222	.0763
	6	0	1.3715	1.2306	.9238	.5784	.2106
7	0	7	1.6502	1.4859	1.1278	.7230	.2779
	1	6	.3715	.5316	.5262	.3601	.1309
	2	5	−.9072	−.4220	−.0696	.0188	.0138
	3	4	−2.1637	−1.3311	−.5815	−.2272	−.0532
	4	3	−2.1637	−1.3311	−.5815	−.2272	−.0532
	5	2	−.9072	−.4220	−.0696	.0188	.0138
	6	1	.3715	.5316	.5262	.3601	.1309
	7	0	1.6502	1.4859	1.1278	.7230	.2779

TABLE 14

c_1

s	s_1	s_2	θ_1				
			.05	.10	.20	.30	.40
2	0	2	.1367	.1042	.0555	.0238	.0058
	1	1	−.1038	−.0840	−.0492	−.0226	−.0058
	2	0	.1367	.1042	.0555	.0238	.0058
3	0	3	.1852	.1392	.0728	.0309	.0075
	1	2	−.0476	−.0380	−.0218	−.0098	−.0025
	2	1	−.0476	−.0380	−.0218	−.0098	−.0025
	3	0	.1852	.1392	.0728	.0309	.0075
4	0	4	.1991	.1447	.0719	.0295	.0071
	1	3	−.0186	−.0117	−.0037	−.0007	0
	2	2	−.0270	−.0245	−.0168	−.0084	−.0023
	3	1	−.0186	−.0117	−.0037	−.0007	0
	4	0	.1991	.1447	.0719	.0295	.0071
5	0	5	.1987	.1382	.0640	.0249	.0058
	1	4	−.0049	.0007	.0047	.0034	.0011
	2	3	−.0133	−.0120	−.0082	−.0041	−.0011
	3	2	−.0133	−.0120	−.0082	−.0041	−.0011
	4	1	−.0049	.0007	.0047	.0034	.0011
	5	0	.1987	.1382	.0640	.0249	.0058
6	0	6	.1921	.1273	.0542	.0197	.0043
	1	5	.0016	.0062	.0077	.0046	.0013
	2	4	−.0064	−.0054	−.0030	−.0012	−.0003
	3	3	−.0068	−.0065	−.0051	−.0028	−.0008
	4	2	−.0064	−.0054	−.0030	−.0012	−.0003
	5	1	.0016	.0062	.0077	.0046	.0013
	6	0	.1921	.1273	.0542	.0197	.0043
7	0	7	.1831	.1153	.0447	.0149	.0031
	1	6	.0046	.0083	.0081	.0044	.0012
	2	5	−.0030	−.0021	−.0005	.0002	.0001
	3	4	−.0034	−.0032	−.0025	−.0014	−.0004
	4	3	−.0034	−.0032	−.0025	−.0014	−.0004
	5	2	−.0030	−.0021	−.0005	.0002	.0001
	6	1	.0046	.0083	.0081	.0044	.0012
	7	0	.1831	.1153	.0447	.0149	.0031

TABLE 11

z_2

s	a	b	c	d	θ_1				
					.05	.10	.20	.30	.40
2	2	0	0	0	.0374	.0298	.0170	.0077	.0019
	1	1	0	0	−.1367	−.1042	−.0555	−.0238	−.0058
	1	0	1	0	−.0410	−.0320	−.0177	−.0078	−.0019
	1	0	0	1	.1038	.0840	.0492	.0226	.0058
	0	2	0	0	.2577	.2148	.1335	.0645	.0170
	0	1	1	0	.1038	.0840	.0492	.0226	.0058
	0	1	0	1	−.7212	−.4437	−.1938	−.0757	−.0177
	0	0	2	0	.0374	.0298	.0170	.0077	.0019
	0	0	1	1	−.1367	−.1042	−.0555	−.0238	−.0058
	0	0	0	2	.2577	.2148	.1335	.0645	.0170
3	3	0	0	0	.1038	.0840	.0492	.0226	.0058
	2	1	0	0	−.2596	−.1908	−.0969	−.0404	−.0098
	2	0	1	0	−.0410	−.0320	−.0177	−.0078	−.0019
	2	0	0	1	.2122	.1754	.1072	.0509	.0133
	1	2	0	0	.1038	.0840	.0492	.0226	.0058
	1	1	1	0	−.0410	−.0320	−.0177	−.0078	−.0019
	1	1	0	1	−.7212	−.4437	−.1938	−.0757	−.0177
	1	0	2	0	−.0410	−.0320	−.0177	−.0078	−.0019
	1	0	1	1	−.0410	−.0320	−.0177	−.0078	−.0019
	1	0	0	2	.3711	.3153	.2041	.1027	.0280
	0	3	0	0	.5353	.4654	.3181	.1703	.0492
	0	2	1	0	.3711	.3153	.2041	.1027	.0280
	0	2	0	1	−.7212	−.4437	−.1938	−.0757	−.0177
	0	1	2	0	.2122	.1754	.1072	.0509	.0133
	0	1	1	1	−.7212	−.4437	−.1938	−.0757	−.0177
	0	1	0	2	−.7212	−.4437	−.1938	−.0757	−.0177
	0	0	3	0	.1038	.0840	.0492	.0226	.0058
	0	0	2	1	−.2596	−.1908	−.0969	−.0404	−.0098
	0	0	1	2	.1038	.0840	.0492	.0226	.0058
	0	0	0	3	.5353	.4654	.3181	.1703	.0492
4	4	0	0	0	.1898	.1559	.0940	.0441	.0114
	3	1	0	0	−.3608	−.2532	−.1219	−.0494	−.0118
	3	0	1	0	−.0035	−.0022	−.0007	−.0001	0
	3	0	0	1	.3231	.2715	.1717	.0843	.0226
	2	2	0	0	−.0492	−.0442	−.0295	−.0144	−.0038
	2	1	1	0	−.1776	−.1362	−.0732	−.0316	−.0078
	2	1	0	1	−.6838	−.4139	−.1768	−.0681	−.0158
	2	0	2	0	−.0819	−.0641	−.0355	−.0156	−.0039
	2	0	1	1	.0628	.0519	.0315	.0148	.0038
	2	0	0	2	.4847	.4166	.2775	.1442	.0406

TABLE 11.—*Continued*

s	a	b	c	d	θ_1				
					.05	.10	.20	.30	.40
	1	3	0	0	.3804	.3311	.2245	.1178	.0332
	1	2	1	0	.2167	.1828	.1158	.0567	.0151
	1	2	0	1	−.8579	−.5479	−.2493	−.0995	−.0236
	1	1	2	0	.0628	.0519	.0315	.0148	.0038
	1	1	1	1	−.7622	−.4757	−.2115	−.0835	−.0197
	1	1	0	2	−.6174	−.3597	−.1446	−.0532	−.0120
	1	0	3	0	−.0035	−.0022	−.0007	−.0001	0
	1	0	2	1	−.1776	−.1362	−.0732	−.0316	−.0078
	1	0	1	2	.2167	.1828	.1158	.0567	.0151
	1	0	0	3	.6492	.5678	.3950	.2171	.0647
	0	4	0	0	.8140	.7201	.5171	.2979	.0940
	0	3	1	0	.6492	.5678	.3950	.2171	.0647
	0	3	0	1	−.4636	−.2289	−.0603	−.0113	−.0007
	0	2	2	0	.4847	.4166	.2775	.1442	.0406
	0	2	1	1	−.6174	−.3597	−.1446	−.0532	−.0120
	0	2	0	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	0	1	3	0	.3231	.2715	.1717	.0843	.0226
	0	1	2	1	−.6838	−.4139	−.1768	−.0681	−.0158
	0	1	1	2	−.8579	−.5479	−.2493	−.0995	−.0236
	0	1	0	3	−.4636	−.2289	−.0603	−.0113	−.0007
	0	0	4	0	.1898	.1559	.0940	.0441	.0114
	0	0	3	1	−.3608	−.2532	−.1219	−.0494	−.0118
	0	0	2	2	−.0492	−.0442	−.0295	−.0144	−.0038
	0	0	1	3	.3804	.3311	.2245	.1178	.0332
	0	0	0	4	.8140	.7201	.5171	.2979	.0940
5	5	0	0	0	.2879	.2396	.1486	.0716	.0189
	4	1	0	0	−.4307	−.2859	−.1294	−.0507	−.0118
	4	0	1	0	.0628	.0519	.0315	.0148	.0038
	4	0	0	1	.4354	.3703	.2407	.1219	.0335
	3	2	0	0	−.2006	−.1678	−.1004	−.0458	−.0116
	3	1	1	0	−.3006	−.2229	−.1146	−.0482	−.0117
	3	1	0	1	−.6174	−.3597	−.1446	−.0532	−.0120
	3	0	2	0	−.0819	−.0641	−.0355	−.0156	−.0039
	3	0	1	1	.1712	.1434	.0895	.0431	.0114
	3	0	0	2	.5985	.5185	.3527	.1886	.0546
	2	3	0	0	.2256	.1972	.1325	.0679	.0187
	2	2	1	0	.0628	.0519	.0315	.0148	.0038
	2	2	0	1	−.9809	−.6345	−.2907	−.1161	−.0275
	2	1	2	0	−.0819	−.0641	−.0355	−.0156	−.0039
	2	1	1	1	−.7622	−.4757	−.2115	−.0835	−.0197

TABLE 11.—*Concluded*

s	a	b	c	d	θ_1				
					.05	.10	.20	.30	.40
	2	1	0	2	−.5091	−.2683	−.0866	−.0248	−.0044
	2	0	3	0	−.0819	−.0641	−.0355	−.0156	−.0039
	2	0	2	1	−.0819	−.0641	−.0355	−.0156	−.0039
	2	0	1	2	.3301	.2832	.1864	.0949	.0261
	2	0	0	3	.7631	.6703	.4727	.2656	.0814
	1	4	0	0	.6591	.5855	.4211	.2401	.0741
	1	3	1	0	.4943	.4333	.3003	.1625	.0473
	1	3	0	1	−.6174	−.3597	−.1446	−.0532	−.0120
	1	2	2	0	.3301	.2832	.1864	.0949	.0261
	1	2	1	1	−.7622	−.4757	−.2115	−.0835	−.0197
	1	2	0	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	1	1	3	0	.1712	.1434	.0895	.0431	.0114
	1	1	2	1	−.7622	−.4757	−.2115	−.0835	−.0197
	1	1	1	2	−.7622	−.4757	−.2115	−.0835	−.0197
	1	1	0	3	−.3502	−.1284	.0103	.0269	.0103
	1	0	4	0	.0628	.0519	.0315	.0148	.0038
	1	0	3	1	−.3006	−.2229	−.1146	−.0482	−.0117
	1	0	2	2	.0628	.0519	.0315	.0148	.0038
	1	0	1	3	.4943	.4333	.3003	.1625	.0473
	1	0	0	4	.9279	.8228	.5958	.3489	.1130
	0	5	0	0	1.0927	.9753	.7200	.4358	.1486
	0	4	1	0	.9279	.8228	.5958	.3489	.1130
	0	4	0	1	−.1860	.0217	.1242	.0945	.0315
	0	3	2	0	.7631	.6703	.4727	.2656	.0814
	0	3	1	1	−.3502	−.1284	.0103	.0269	.0103
	0	3	0	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	0	2	3	0	.5985	.5185	.3527	.1886	.0546
	0	2	2	1	−.5091	−.2683	−.0866	−.0248	−.0044
	0	2	1	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	0	2	0	3	−1.4425	−.8874	−.3876	−.1514	−.0355
	0	1	4	0	.4354	.3703	.2407	.1219	.0335
	0	1	3	1	−.6174	−.3597	−.1446	−.0532	−.0120
	0	1	2	2	−.9809	−.6345	−.2907	−.1161	−.0275
	0	1	1	3	−.6174	−.3597	−.1446	−.0532	−.0120
	0	1	0	4	−.1860	.0217	.1242	.0945	.0315
	0	0	5	0	.2879	.2396	.1486	.0716	.0189
	0	0	4	1	−.4307	−.2859	−.1294	−.0507	−.0018
	0	0	3	2	−.2006	−.1678	−.1004	−.0458	−.0116
	0	0	2	3	.2256	.1972	.1325	.0679	.0187
	0	0	1	4	.6591	.5855	.4211	.2401	.0741
	0	0	0	5	1.0927	.9753	.7200	.4358	.1486

TABLE 12

z_2

s	a	b+c	d	θ_1				
				.05	.10	.20	.30	.40
2	2	0	0	.0120	.0090	.0045	.0018	.0004
	1	1	0	−.0382	−.0281	−.0139	−.0056	−.0013
	1	0	1	.0979	.0747	.0392	.0164	.0039
	0	2	0	.0979	.0747	.0392	.0164	.0039
	0	1	1	−.6174	−.3597	−.1446	−.0532	−.0120
	0	0	2	.5154	.4297	.2671	.1289	.0341
3	3	0	0	.0373	.0277	.0139	.0056	.0013
	2	1	0	−.0740	−.0528	−.0249	−.0096	−.0022
	2	0	1	.1993	.1543	.0824	.0346	.0083
	1	2	0	.0542	.0386	.0175	.0063	.0014
	1	1	1	−.5782	−.3270	−.1244	−.0435	−.0094
	1	0	2	.6252	.5235	.3273	.1582	.0417
	0	3	0	.2076	.1680	.0984	.0451	.0115
	0	2	1	−.7622	−.4757	−.2115	−.0835	−.0197
	0	1	2	−.3502	−.1284	.0103	.0269	.0103
	0	0	3	1.0706	.9308	.6361	.3405	.0984
4	4	0	0	.0763	.0568	.0283	.0114	.0026
	3	1	0	−.1064	−.0732	−.0325	−.0121	−.0027
	3	0	1	.3034	.2378	.1293	.0547	.0131
	2	2	0	.0108	.0034	−.0026	−.0025	−.0008
	2	1	1	−.5261	−.2848	−.0995	−.0319	−.0065
	2	0	2	.7353	.6180	.3891	.1888	.0498
	1	3	0	.1632	.1298	.0727	.0317	.0078
	1	2	1	−.7877	−.4859	−.2092	−.0801	−.0185
	1	1	2	−.2462	−.0439	.0608	.0509	.0166
	1	0	3	1.1811	1.0270	.7031	.3775	.1093
	0	4	0	.3187	.2657	.1676	.0831	.0225
	0	3	1	−.6937	−.4465	−.2188	−.0938	−.0232
	0	2	2	−1.0746	−.5775	−.1905	−.0539	−.0092
	0	1	3	.1856	.3389	.3347	.2058	.0640
	0	0	4	1.6280	1.4403	1.0343	.5958	.1880
5	5	0	0	.1286	.0961	.0480	.0191	.0044
	4	1	0	−.1343	−.0884	−.0365	−.0128	−.0027
	4	0	1	.4092	.3245	.1795	.0766	.0183
	3	2	0	−.0322	−.0307	−.0208	−.0100	−.0026
	3	1	1	−.4620	−.2335	−.0698	−.0185	−.0031
	3	0	2	.8456	.7130	.4522	.2206	.0582
	2	3	0	.1189	.0920	.0478	.0193	.0044
	2	2	1	−.8075	−.4900	−.2029	−.0750	−.0169
	2	1	2	−.1404	.0435	.1142	.0764	.0233
	2	0	3	1.2917	1.1233	.7706	.4152	.1206

TABLE 12.—*Continued*

s	a	b+c	d	θ_1				
				.05	.10	.20	.30	.40
	1	4	0	.2741	.2268	.1397	.0671	.0177
	1	3	1	−.7331	−.4741	−.2286	−.0958	−.0233
	1	2	2	−1.0107	−.5235	−.1555	−.0361	−.0042
	1	1	3	.2959	.4340	.3987	.2396	.0737
	1	0	4	1.7386	1.5367	1.1031	.6366	.2015
	0	5	0	.4301	.3649	.2422	.1278	.0366
	0	4	1	−.5931	−.3728	−.1905	−.0879	−.0228
	0	3	2	−1.3547	−.7977	−.3166	−.1123	−.0244
	0	2	3	−.6897	−.2365	.0540	.0851	.0334
	0	1	4	.7420	.8444	.7200	.4434	.1445
	0	0	5	2.1855	1.9507	1.4400	.8717	.2972
6	6	0	0	.1932	.1451	.0728	.0289	.0066
	5	1	0	−.1561	−.0971	−.0364	−.0117	−.0023
	5	0	1	.5165	.4135	.2327	.1002	.0240
	4	2	0	−.0746	−.0633	−.0368	−.0160	−.0039
	4	1	1	−.3875	−.1738	−.0356	−.0031	.0008
	4	0	2	.9559	.8084	.5163	.2536	.0671
	3	3	0	.0747	.0545	.0240	.0078	.0015
	3	2	1	−.8201	−.4867	−.1923	−.0681	−.0148
	3	1	2	−.0331	.1330	.1701	.1034	.0305
	3	0	3	1.4023	1.2196	.8383	.4535	.1322
	2	4	0	.2296	.1882	.1123	.0518	.0132
	2	3	1	−.7718	−.4999	−.2360	−.0964	−.0231
	2	2	2	−.9364	−.4615	−.1163	−.0165	.0012
	2	1	3	.4062	.5295	.4636	.2744	.0838
	2	0	4	1.8492	1.6332	1.1720	.6778	.2154
	1	5	0	.3855	.3257	.2128	.1098	.0308
	1	4	1	−.6342	−.4049	−.2072	−.0942	−.0242
	1	3	2	−1.3672	−.7915	−.3002	−.1008	−.0208
	1	2	3	−.5826	−.1466	.1117	.1147	.0419
	1	1	4	.8525	.9408	.7880	.4827	.1571
	1	0	5	2.2961	2.0472	1.5093	.9143	.3129
	0	6	0	.5418	.4651	.3200	.1776	.0536
	0	5	1	−.4891	−.2888	−.1441	−.0693	−.0188
	0	4	2	−1.3191	−.8168	−.3702	−.1488	−.0353
	0	3	3	−1.4833	−.7448	−.1870	−.0182	.0070
	0	2	4	−.1435	.2505	.4115	.2985	.1043
	0	1	5	1.2994	1.3544	1.1224	.7109	.2465
	0	0	6	2.7430	2.4612	1.8476	1.1568	.4212

TABLE 12.—*Concluded*

s	a	b+c	d	θ_1				
				.05	.10	.20	.30	.40
7	7	0	0	.2684	.2032	.1028	.0408	.0093
	6	1	0	−.1703	−.0981	−.0319	−.0087	−.0015
	6	0	1	.6247	.5044	.2884	.1255	.0302
	5	2	0	−.1162	−.0939	−.0504	−.0206	−.0048
	5	1	1	−.3043	−.1068	.0029	.0141	.0051
	5	0	2	1.0663	.9041	.5814	.2876	.0764
	4	3	0	.0306	.0175	.0014	−.0025	−.0011
	4	2	1	−.8237	−.4752	−.1774	−.0594	−.0124
	4	1	2	.0750	.2243	.2281	.1318	.0380
	4	0	3	1.5129	1.3160	.9064	.4925	.1442
	3	4	0	.1852	.1497	.0855	.0374	.0091
	3	3	1	−.8095	−.5233	−.2406	−.0954	−.0223
	3	2	2	−.8534	−.3925	−.0732	.0048	.0070
	3	1	3	.5166	.6252	.5293	.3101	.0942
	3	0	4	1.9597	1.7297	1.2410	.7193	.2296
	2	5	0	.3409	.2866	.1838	.0925	.0253
	2	4	1	−.6751	−.4365	−.2225	−.0994	−.0252
	2	3	2	−1.3708	−.7769	−.2793	−.0875	−.0167
	2	2	3	−.4745	−.0551	.1712	.1455	.0508
	2	1	4	.9631	1.0372	.8563	.5224	.1700
	2	0	5	2.4067	2.1437	1.5786	.9571	.3287
	1	6	0	.4970	.4254	.2896	.1581	.0469
	1	5	1	−.5303	−.3219	−.1642	−.0790	−.0213
	1	4	2	−1.3565	−.8385	−.3696	−.1432	−.0331
	1	3	3	−1.4009	−.6749	−.1409	.0063	.0142
	1	2	4	−.0331	.3463	.4777	.3355	.1158
	1	1	5	1.4100	1.4509	1.1914	.7528	.2614
	1	0	6	2.8536	2.5577	1.9170	1.2003	.4384
	0	7	0	.6537	.5660	.4004	.2315	.0732
	0	6	1	−.3846	−.2024	−.0886	−.0413	−.0112
	0	5	2	−1.2229	−.7576	−.3720	−.1660	−.0422
	0	4	3	−1.9107	−1.0674	−.3649	−.1010	−.0153
	0	3	4	−1.0240	−.3345	.1158	.1639	.0677
	0	2	5	.4134	.7584	.8062	.5536	.1985
	0	1	6	1.8569	1.8649	1.5291	.9923	.3651
	0	0	7	3.3005	2.9718	2.2557	1.4460	.5559

TABLE 13

n	c_1					c_2					c_3				
	θ_1					θ_1					θ_1				
	.05	.10	.20	.30	.40	.05	.10	.20	.30	.40	.05	.10	.20	.30	.40
2	−.0374	−.0298	−.0170	−.0077	−.0019	−.0164	−.0130	−.0074	−.0033	−.0008	−.0197	−.0147	−.0075	−.0031	−.0007
3	−.0210	−.0167	−.0095	−.0042	−.0011	−.0121	−.0096	−.0054	−.0024	−.0006	−.0203	−.0150	−.0075	−.0030	−.0007
4	−.0105	−.0081	−.0044	−.0019	−.0005	−.0073	−.0057	−.0032	−.0014	−.0004	−.0175	−.0128	−.0062	−.0025	−.0006
5	−.0051	−.0038	−.0019	−.0008	−.0002	−.0041	−.0032	−.0018	−.0008	−.0002	−.0143	−.0103	−.0049	−.0019	−.0004
6	−.0025	−.0017	−.0008	−.0003	−.0001	−.0022	−.0017	−.0009	−.0004	−.0001	−.0113	−.0080	−.0037	−.0014	−.0003
7	−.0012	−.0008	−.0003	−.0001	0	−.0011	−.0009	−.0005	−.0002	0	−.0087	−.0061	−.0027	−.0010	−.0002
8	−.0006	−.0004	−.0001	0	0	−.0006	−.0004	−.0002	−.0001	0	−.0067	−.0046	−.0020	−.0007	−.0002
9	−.0003	−.0002	−.0001	0	0	−.0003	−.0002	−.0001	0	0	−.0050	−.0034	−.0014	−.0005	−.0001
10	−.0001	−.0001	0	0	0	−.0001	−.0001	−.0001	0	0	−.0038	−.0025	−.0010	−.0003	−.0001
11	−.0001	0	0	0	0	−.0001	−.0001	0	0	0	−.0028	−.0018	−.0007	−.0002	0
12	0	0	0	0	0	0	0	0	0	0	−.0021	−.0013	−.0005	−.0002	0
13											−.0016	−.0010	−.0003	−.0001	0
14											−.0011	−.0007	−.0002	−.0001	0
15											−.0008	−.0005	−.0002	0	0

not affect the validity of a sequential test, but his proof of the optimum character of the sequential test does not cover dependent observations. I suspect, but have not been able to prove, that the sequential probability ratio test is optimum for this class of dependence also.

The ease and exactness with which probability ratio scores may be combined is particularly important when the data are of mixed known and unknown phase, since the alternative u score theory provides only a rough approximation in small samples (Finney, 1943; Smith, 1953). This is a critical point, not only for human pedigrees, but especially in laboratory vertebrates where linkage studies are of secondary interest and the material on any particular pair of loci is usually heterogeneous and small.

16. INSTRUCTIONS FOR ANALYSIS

Although the simplicity of the sequential probability ratio test allows the investigator to modify his methods to fit particular situations, it may be useful to set down here instructions for the routine case of unrelated families, tested parents, known parental genotypes, and unknown phase.

Step 1. Define the method of selection. This comprehends both ascertainment of families and rejection of some kinds of ascertained families. Usually, families with untested parents or of doubtful mating type will be rejected; otherwise, cf. §§12–13. For each factor selection may be complete, truncate, or arbitrary (§7). With respect to the two factors in a linkage test, there are three important methods of selection:

(i) Complete selection of one or both factors.

(ii) Truncate selection of both factors.

(iii) Arbitrary selection of one factor (G), truncate selection of the other (T).

Step 2. Choose the alternative hypothesis (cf. §15). If the amount of data that can be obtained with "reasonable" effort is likely to be small, choose $\theta_1 = .05$ or .10; if a moderately large amount of data is hoped for, choose $\theta_1 = .20$ or .30; if an extraordinarily large amount is anticipated, take $\theta_1 = .40$. Usually, $\log B = -2$ and $\log A = 3$ are appropriate choices for the other parameters of the test.

Step 3. Classify the mating type of each family according to tables 4–8, and distribute the children among classes a, b, c, d, $\cdots$. In these tables, G_1, G_2 and T_1, T_2 denote factors without dominance or rare "dominants", while G, g and T, t are factors showing simple dominant-recessive relationships.

Step 4. Determine the score for each family from tables 10–18, or compute directly, using common logarithms. The following outline may be helpful in performing the above steps.

Classification of matings, methods of selection, and scores (z)

I. Double backcross, and single backcross with no dominance in the intercross factor.

(i) Complete selection of either factor	z_1
(ii) Truncate selection of both factors	$z_1 + c_1$
(iii) Arbitrary-truncate selection	$z_1 + e_1$

TABLE 15

s	s_1	s_2	c_2					d_2					e_2				
			θ_1					θ_1					θ_1				
			.05	.10	.20	.30	.40	.05	.10	.20	.30	.40	.05	.10	.20	.30	.40
2	0	2	.1367	.1042	.0555	.0238	.0058	.0534	.0416	.0229	.0100	.0025	.1708	.1200	.0562	.0220	.0051
	1	1	−.0374	−.0298	−.0170	−.0077	−.0019	−.0476	−.0380	−.0218	−.0098	−.0025	−.0447	−.0336	−.0172	−.0071	−.0017
	2	0	.0132	.0104	.0058	.0026	.0006	.0534	.0416	.0229	.0100	.0025	.0160	.0118	.0059	.0024	.0006
3	0	3	.1852	.1392	.0728	.0309	.0075	.0953	.0735	.0398	.0172	.0042	.2575	.1835	.0892	.0362	.0086
	1	2	.0171	.0134	.0075	.0033	.0008	−.0276	−.0220	−.0125	−.0056	−.0014	.0443	.0305	.0132	.0047	.0010
	2	1	−.0271	−.0215	−.0122	−.0055	−.0014	−.0276	−.0220	−.0125	−.0056	−.0014	−.0470	−.0348	−.0173	−.0069	−.0016
	3	0	.0171	.0134	.0075	.0033	.0008	.0953	.0735	.0398	.0172	.0042	.0294	.0214	.0104	.0041	.0010
4	0	4	.1991	.1447	.0719	.0295	.0071	.1271	.0968	.0515	.0221	.0054	.3022	.2150	.1068	.0446	.0108
	1	3	.0426	.0344	.0201	.0091	.0023	−.0016	−.0010	−.0003	−.0001	0	.1030	.0764	.0383	.0155	.0036
	2	2	−.0031	−.0028	−.0019	−.0009	−.0003	−.0332	−.0267	−.0155	−.0070	−.0018	.0051	.0016	−.0012	−.0011	−.0004
	3	1	−.0160	−.0126	−.0070	−.0031	−.0008	−.0016	−.0010	−.0003	−.0001	0	−.0416	−.0301	−.0143	−.0055	−.0012
	4	0	.0162	.0127	.0071	.0031	.0008	.1271	.0968	.0515	.0221	.0054	.0404	.0290	.0138	.0054	.0012
5	0	5	.1987	.1382	.0640	.0249	.0058	.1504	.1130	.0590	.0249	.0061	.3247	.2290	.1148	.0489	.0120
	1	4	.0530	.0419	.0236	.0105	.0026	.0216	.0175	.0103	.0047	.0012	.1389	.1043	.0545	.0232	.0057
	2	3	.0097	.0081	.0050	.0024	.0006	−.0226	−.0182	−.0105	−.0047	−.0012	.0448	.0331	.0160	.0062	.0014
	3	2	−.0052	−.0045	−.0029	−.0014	−.0004	−.0226	−.0182	−.0105	−.0047	−.0012	−.0096	−.0091	−.0063	−.0031	−.0008
	4	1	−.0087	−.0067	−.0036	−.0015	−.0004	.0216	.0175	.0103	.0047	.0012	−.0345	−.0242	−.0108	−.0039	−.0008
	5	0	.0134	.0104	.0058	.0025	.0006	.1504	.1130	.0590	.0249	.0061	.0493	.0348	.0161	.0061	.0014
6	0	6	.1921	.1273	.0542	.0197	.0043	.1667	.1235	.0630	.0263	.0063	.3347	.2330	.1168	.0503	.0125
	1	5	.0564	.0429	.0226	.0096	.0023	.0402	.0322	.0184	.0083	.0021	.1608	.1202	.0639	.0279	.0070
	2	4	.0154	.0128	.0078	.0036	.0009	−.0091	−.0071	−.0039	−.0017	−.0004	.0710	.0542	.0287	.0121	.0029
	3	3	.0012	.0011	.0007	.0003	.0001	−.0226	−.0183	−.0107	−.0049	−.0012	.0180	.0128	.0054	.0017	.0003
	4	2	−.0038	−.0033	−.0021	−.0010	−.0003	−.0091	−.0071	−.0039	−.0017	−.0004	−.0146	−.0126	−.0076	−.0034	−.0008
	5	1	−.0045	−.0033	−.0017	−.0007	−.0002	.0402	.0322	.0184	.0083	.0021	−.0275	−.0184	−.0075	−.0025	−.0005
	6	0	.0103	.0080	.0043	.0019	.0005	.1667	.1235	.0630	.0263	.0063	.0562	.0390	.0175	.0065	.0015

7	0	7	.1831	.1153	.0447	.0149	.0031	.1776	.1294	.0645	.0265	0063	.3371	.2314	.1153	.0499	.0125
	1	6	.0564	.0411	.0200	.0079	.0018	.0545	.0431	.0242	.0107	.0027	.1738	.1286	.0686	.0304	.0077
	2	5	.0177	.0142	.0082	.0037	.0009	.0035	.0031	.0022	.0011	.0003	.0881	.0675	.0368	.0162	.0040
	3	4	.0042	.0037	.0024	.0012	.0003	−.0162	−.0131	−.0076	−.0035	−.0009	.0371	.0285	.0148	.0061	.0014
	4	3	−.0006	−.0005	−.0003	−.0002	0	−.0162	−.0131	−.0076	−.0035	−.0009	.0049	.0028	.0002	−.0004	−.0002
	5	2	−.0023	−.0020	−.0013	−.0006	−.0002	.0035	.0031	.0022	.0011	.0003	−.0154	−.0128	−.0073	−.0031	−.0007
	6	1	−.0022	−.0015	−.0007	−.0002	0	.0545	.0431	.0242	.0107	.0027	−.0212	−.0133	−.0047	−.0013	−.0002
	7	0	.0075	.0058	.0031	.0013	.0003	.1776	.1294	.0645	.0265	.0063	.0613	.0419	.0182	.0066	.0015

TABLE 18.—LOD SCORES FOR INDIVIDUAL PROGENY WHEN THE PARENTAL PHASE IS KNOWN

$\frac{f(y;\theta_1)}{f(y;\frac{1}{2})}$	θ_1				
	.05	.10	.20	.30	.40
$2\theta_1$	−1.0000	−.6990	−.3979	−.2218	−.0969
$2(1-\theta_1)$	.2788	.2553	.2041	.1461	.0792
$2(2-\theta_1)/3$	.1139	.1027	.0792	.0544	.0280
$2(1+\theta_1)/3$	−.1549	−.1347	−.0969	−.0621	−.0300
$4(3-2\theta_1+\theta_1^2)/9$	.1106	.0965	.0694	.0440	.0207
$4(2+\theta_1-\theta_1^2)/9$	−.0410	−.0320	−.0177	−.0078	−.0019
$4(1-\theta_1+\theta_1^2)/3$	.1038	.0840	.0492	.0226	.0058
$4(2+\theta_1^2)/9$	−.0506	−.0490	−.0426	−.0320	−.0177
$2(1+2\theta_1-2\theta_1^2)/3$	−.1367	−.1042	−.0555	−.0238	−.0058
$2(1-2\theta_1+2\theta_1^2)$	.2577	.2148	.1335	.0645	.0170

II. Single backcross with dominance in the intercross factor.
  (i) Complete selection of either factor z_2
  (ii) Truncate selection of both factors $z_2 + c_2$
  (iii) Arbitrary selection of intercross factor, truncate selection of backcross factor $z_2 + e_2$
  (iv) Arbitrary selection of backcross factor, truncate selection of intercross factor $z_2 + d_2$

III. Double intercross with dominance in both factors
  (i) Complete selection of either factor z_3
  (ii) Truncate selection of both factors $z_3 + c_3$
  (iii) Arbitrary-truncate selection $z_3 + e_3$

IV. Double intercross with dominance in one factor
  (i) Complete selection of either factor z_4
  (ii) Arbitrary selection of factor with no dominance, truncate selection of dominant factor $z_4 + e_4$

V. Double intercross with no dominance in either factor
  (i) Complete selection of either factor z_5

Step 5. Accumulate the family scores (z). If $\sum z \leq \log B$, conclude that the frequency of recombination θ is significantly greater than θ_1 on the assumptions of §1. If $\sum z \geq \log A$, conclude that θ is significantly less than 1/2. Review the data and assumptions before deciding that true linkage is present. If $\log B < \sum z < \log A$, suspend judgment about linkage until further data lead to a decision. More data can also be used to estimate θ, after linkage has been detected, or to make a further test for linkage in the range $\theta_1 < \theta < 1/2$, if that seems advisable.

The following examples illustrate the scoring procedure.

Case 1. A mating of type GT × gt gives 2GT, 2Gt, and 1gt progeny. This is a double backcross (mating 1) with s = 5, a + d = 3. The score for complete selection is z_1 (table 10). For truncate selection of both factors, add the correction factor c_1 (table 13), and for truncate selection of the T factor but arbitrary selection of the G factor (which shows 4G:1g) add e_1 with $s_1 = 4$, $s_2 = 1$ (table 14). For $\theta_1 = .20$, we find $z_1 = -.3876$, $z_1 + c_1 = -.3895$, and $z_1 + e_1 = -.3829$.

Case 2. A mating of type GT × Gt gives 5GT, 2gT, 3Gt, and 1gt progeny. This is a single backcross (mating 9) with s = 11, a = 5, b = 2, c = 3, and d = 1. Families of this size are not given in table 11, but the score may quickly be obtained by factoring the expression for z_2 which is

$$z_2 = \log \frac{2^{10}}{3^8} [(2 - \theta_1)^5\theta_1^2 (1 + \theta_1)^3(1 - \theta_1) + (1 + \theta_1)^5 (1 - \theta_1)^2(2 - \theta_1)^3\theta_1]$$

$$= 3 \log [2(2 - \theta_1)/3] + 3 \log [2(1 + \theta_1)/3] + \log 2\theta_1 + \log 2(1 - \theta_1)$$

$$+ \log \frac{2^2}{3^2} [(2 - \theta_1)^2\theta_1 + (1 + \theta_1)^2(1 - \theta_1)].$$

The first four terms correspond to progeny of known parental phase (table 18), the last term to a single backcross family with s = 3, a = 2, b = 1, c = d = 0. For $\theta_1 = .20$, we find

$$z_2 = 3(.0792) + 3(-.0969) + (-.3979) + .2041 + (-.0969) = -.3438.$$

The corresponding scores for incomplete selection are $z_2 + c_2 = -.3438$ and $z_2 + e_2 = -.3439$. Here, as is usual in large families, the corrections for incomplete selection are negligible.

17. SUMMARY

The sequential probability ratio test for linkage detection in man is simple, exact and efficient. The basic assumptions of the linkage test are discussed, and criteria are developed for the choice of parameters in the sequential test. For the case of double backcross sib-pairs, the sequential tests considered here require less than 1/3 as many observations for a given risk of error as the Fisher-Finney u score method and about 1/5 as many observations as the Haldane-Smith nonsequential probability ratio test. Formulae for "lod" scores are given for a variety of mating types and methods of selection, and the research worker should have no difficulty extending the formulae to novel cases as they arise. The optimum property of the sequential probability ratio test holds for mixed data, the combination of which is easy and exact. Examples and tables of scores are given for the most important mating types.

The work for this paper was done under the direction of Dr. J. F. Crow, to whom the author is indebted for many stimulating discussions and constant encouragement. Drs. E. R. Immel, W. J. Schull, and C. A. B. Smith read the preliminary manuscript and offered helpful comments. Thanks are also due to the Numerical Analysis Laboratory of the University of Wisconsin, and especially to Mr. William Graebel, for assistance in computing the tables of linkage scores.

REFERENCES

BAILEY, N. T. J. 1951. A classification of methods of ascertainment and analysis in estimating the frequencies of recessives in man. *Ann. Eugen.* 16: 223–225.

BERNSTEIN, F. 1931. Zur Grundlegung der Chromosomentheorie der Vererbung beim Menschen mit besondere Berücksichtung der Blutgruppen. *Z. indukt. Abstamm. u. VererbLehre* 57: 113–138.

BRIDGES, C. B., AND K. S. BREHME 1944. The mutants of Drosophila melanogaster. *Carnegie Inst. Wash. Publ.* 552.

BROSS, I. 1952. Sequential medical plans. *Biometrics* 8: 188–205.
CARTER, T. C. 1955. The estimation of total genetical map lengths from linkage test data. *J. Gener.* 53: 21–28.
CREW, F. A. E., AND P. CH. KOLLER 1932. The sex incidence of chiasma frequency and genetical crossing-over in the mouse. *J. Genet.* 26: 359–384.
FINNEY, D. J. 1940. The detection of linkage. *Ann. Eugen.* 10: 171–214.
FINNEY, D. J. 1941a. The detection of linkage. II: Further mating types; scoring of Boyd's data. *Ann. Eugen.* 11: 10–30.
FINNEY, D. J. 1941b. The detection of linkage. III: Incomplete parental testing. *Ann. Eugen.* 11: 115–135.
FINNEY, D. J. 1942. The detection of linkage. VI: The loss of information from incompleteness of parental testing. *Ann. Eugen.* 11: 233–242.
FINNEY, D. J. 1943. The detection of linkage. VII: Combination of data from matings of known and unknown phase. *Ann. Eugen.* 12: 31–43.
FISHER, R. A. 1935. The detection of linkage with "dominant" abnormalities. *Ann. Eugen.* 6: 187–201.
HALDANE, J. B. S. 1934. Methods for the detection of autosomal linkage in man. *Ann. Eugen.* 6: 26–65.
HALDANE, J. B. S. 1946. The cumulants of the distribution of Fisher's "u_{1i}" and "u_{3i}" scores used in the detection and estimation of linkage in man. *Ann. Eugen.* 13: 122–134.
HALDANE, J. B. S., AND C. A. B. SMITH 1947. A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Ann. Eugen.* 14: 10–31.
HARRIS, H. 1948. On sex limitation in human genetics. *Eugen. Rev.* 40: 70–76.
HOGBEN, L. 1934. The detection of linkage in human families. *Proc. Roy. Soc. B* 114: 340–363.
KOSAMBI, D. D. 1944. The estimation of map distances from recombination values. *Ann. Eugen.* 12: 172–175.
MOHR, J. 1954. A study of linkage in man. *Op. Dom. Biol. Hered. Hum. Univ. Hafn.* 33: 1–119.
NEEL, J. V. 1949. The detection of the genetic carriers of hereditary disease. *Amer. J. Hum. Genet.* 1: 19–36.
PENROSE, L. S. 1953. The general purpose sib-pair linkage test. *Ann. Eugen.* 18: 120–124.
RHOADES, M. M. 1950. Meiosis in maize. *J. Hered.* 41: 59–70.
SLIZYNSKI, B. M. 1949. A preliminary pachytene chromosome map of the house mouse. *J. Genet.* 49: 242–245.
SMITH, C. A. B. 1953. The detection of linkage in human genetics. *J. Roy. Stat. Soc. B* 15: 153–192.
WALD, A. 1947. *Sequential Analysis.* New York: Wiley.
WALD, A., AND J. WOLFOWITZ. 1948. Optimum character of the sequential probability ratio test. *Ann. Math. Stat.* 19: 326–339.

BOOK REVIEWS

Dystrophia Musculorum Progressiva: Eine genetische and klinische Untersuchung der Muskeldystrophien

By Prof. P. E. Becker, Tuttlingen. Georg Thieme Verlag, Stuttgart, 1953. Pp. 311, 101 figures. DM 28.50.

This monograph reports in detail a genetic and clinical investigation of progressive muscular dystrophy in the province of Baden, Germany. The material was collected during the period from 1938 to 1940, and includes information on 259 cases, 162 of these being alive at the time of investigation.

Becker's results confirm those of previous investigators in finding two major types of progressive muscular dystrophy that are genetically distinct. Nine kindreds are described in which the pectoral girdle type is transmitted as a dominant characteristic. Eleven apparently isolated cases of the pectoral girdle type were found, some probably representing new mutations, but the writer suggests that extrinsic factors, perhaps trauma, may be of importance in some instances. The pelvic girdle or childhood type of muscular dystrophy was found in 64 sibships, 37 of these containing only one affected individual. Evidence is presented showing that this group contains families in which the disease is transmitted as a sex-linked recessive characteristic and families in which those affected are homozygous for an autosomal recessive gene. The investigator is unable to distinguish the sex-linked and autosomal types of childhood muscular dystrophy clinically. The prevalence of the two types in the South Baden area as of July, 1939, is estimated at 0.06 per thousand for the pectoral type and 0.05 per thousand for the pelvic type.

Dr. Becker presents his clinical findings and pedigree information in detail. Each kindred is completely described, with pedigree drawings, this section occupying 145 pages of the text. The data are analyzed extensively from both the clinical and genetic viewpoints. Data are included on other disease conditions found in the families studied. The extensive presentation of all available data will make this monograph quite valuable to those doing similar research in other areas. Intensive surveys of limited geographic areas for specific diseases or groups of diseases of genetic origin are of considerable significance to studies of population genetics. This monograph is a very worthwhile addition to the literature in this field. An index is not provided, but the table of contents is quite detailed as to the topics discussed. The bibliography is extensive, occupying 13 pages. The book is attractively printed and bound, and the illustrations, tables and charts are adequate.

C. Nash Herndon, M.D.
Bowman Gray School of Medicine

Genetic Homeostasis

By I. Michael Lerner. New York: John Wiley & Sons, Inc., 1954, pp. 134. $3.25.

The history of genetics, as of other sciences, is characterized by the sporadic appearance of cases of new phenomena. At first the results are uninterpretable; then aspects of similarity emerge ultimately leading to a new synthesis and a fresh impetus for further study.

Perhaps this sequence is nowhere better illustrated than in the developments which led to the concept of genetic homeostasis as described in this book by I. Michael Lerner.

It is necessary at the outset to distinguish between various kinds of homeostasis. *Physiological homeostasis* is "the totality of steady states maintained in an organism through the co-ordination of its complex physiological processes." *Developmental homeostasis* refers to the stabilizing processes embryonic development which tend to eliminate extremes. Similarly, psychological homeostasis and ecological homeostasis, may be defined. In particular, *genetic homeostasis* is "the property of (a mendelian) population to equilibrate its genetic composition and to resist sudden changes." Genetic

homeostasis, as Lerner points out, bridges the gap between individual physiological homeostasis on the one hand and group ecological homeostasis on the other.

The definition of genetic homeostasis raises two questions: what is the evidence for existence of this property in mendelian populations and what mechanisms achieve it. Lerner's book is devoted in about equal measures to discussing these questions.

The evidence is chiefly of three types. First, several selection experiments have resulted in cessation or at least deceleration of progress without a corresponding reduction in genetic variance. It appears that mendelian populations harbor sufficient genetic variability to permit them to respond to artificial selection, but that after a certain limit is reached natural selection opposes the artificial selection, thereby preventing further progress even in the demonstrable presence of a supply of residual genetic variability. Second, numerous cases of phenotypic balance in sparrows, lizards, mice, rabbits, sheep, rats, and other forms suggest that phenotypic deviants are sacrificed to natural selection and that selection favors those organisms which exhibit mean or near mean values of all quantitatively varying traits.

The third type of evidence consists of several cases in Drosophila, poultry, mice, and other forms of the environmental component of variance being larger in homozygotes, smaller in heterozygotes.

The interpretation of these phenomena in genetic, biochemical, and developmental terms provides fascinating opportunities for speculation. Lerner explores several concepts and models of gene action and interrelationship which, if not final, will at least open the way for further experimentation and theory.

The student of human heredity will be especially interested in the concept of genetic homeostasis, since mankind propagates itself by means of a breeding structure which insures the maintenance of maximal or near maximal amounts of heterozygosity.

EARL L. GREEN
Atomic Energy Commission

Man's Capacity to Reproduce the Demography of a Unique Population

By JOSEPH W. EATON and ALBERT J. MAYER, Glencoe, Illinois: The Free Press, 1954, Pp. 59, $2.00.

A HUMAN geneticist might well dream of a population like the Hutterites. They are a vigorous and interesting people. A lot of them are living within short distances of one another. Their families have a median of 9 children. Their non-genetic environment is relatively uniform as Western populations go. They have comparatively excellent knowledge about details of their own family histories. They have a favorable attitude toward scientific and medical investigation. The only disappointment a dreaming medical geneticist might encounter is that they are unusually healthy.

This Free Press edition is a reprint of a paper which appeared in *Human Biology* (Vol. 25, no. 3, September, 1953, pp. 206–264). The original title is "The Social Biology of Very High Fertility among the Hutterites." The subtitles are identical in the two editions.

The Hutterites are an anabaptist sect living (as of 1950) in 93 self-contained colonies in the Dakotas, Montana, Alberta, Saskatchewan and Manitoba. The sect originated in Switzerland and Bohemia in 1528. During the 17th and 18th centuries they suffered severe persecution at the hands of both Catholics and Protestants. In 1762 some members of the sect found sanctuary in Crimea. Between 1874 and 1877 nearly all of the faithful Hutterites, fearful of renewed persecution, moved from Russia to South Dakota. Other Hutterite colonies with slightly different backgrounds are living today in England and Paraguay.

Between 1880 and 1950 the Hutterite population in North America increased over 19 times, from 443 to 8,542. New colonies are formed at a rate which keeps the typical colony size around 100 individuals. 70% of the 443 colony Hutterites listed in the 1880 census had only 5 patronyms and 10 additional surnames account for the other 30%.

The demographic data analyzed by Eaton and Mayer includes records collected by them and their collaborators on 6,796 individuals living in 71 colonies. The data represent a sample of 80% of the total Hutterite population living in 1950 and include information on birthdate, birth order including

stillbirths, sex, marital status, occupation and religious leadership, death dates for the decade 1940–1950, and the place of residence of a small number of Hutterites living outside of a colony.

The reproductive performance of the Hutterites is unique among modern Western populations. The sex ratio is 101♂:100♀. The population is very young—50.6% are under 15 years of age and only 2.3% are over 65. More males than females reach ages over 40. The crude birth rate is 45.9 per 1,000 population. The fertility ratio, that is, the number of children under 5 years per 100 women of ages 15–49, is 96.3. The age specific fertility rate is 391.1 per 1,000 women in the age group 30–34. The 166 women who were between 45 and 54 in 1950 had a mean of 10.6 live births. Nearly everyone marries (only after baptism at the age of 19) and only 3.4% of marriages are childless. All of this reflects the extremely high Hutterite population fecundity: 1 out of each 13 ovulation cycles terminates in a live birth. This estimate of fecundity is minimum because it is not corrected for illness, miscarriages, nor normal separation of consorts.

The crude death rate (1941–1950) was 4.4 per 1,000 population. At the current rate of increase (4.13% per year), the population will double in number in about 16 years.

The Hutterites are an ideal population for study of the interplay of genetic, psychological, social and cultural variables associated with high fertility. Eaton and Mayer plan such a study for the near future. We wish them and their Hutterite collaborators all success in this work.

J. N. Spuhler
Institute of Human Biology
University of Michigan

Human Heredity

By James V. Neel and William J. Schull, Chicago University Press, 1954, Pp. 361.

Human genetics has come a long way since the days when efforts were devoted mainly to the collection of pedigrees of rare conditions and to the fitting (by hook or by crook) of the familial distributions of human traits to Mendelian expectations. Much of the progress has been due to the development of mathematical techniques to deal with the special problems that arise in the analysis of human family data. Until now most of these techniques have remained scattered through a variety of genetic, biometric and medical journals. The present book draws together a number of the more useful statistical methods and discusses their application to current problems in human heredity. Priscilla and Vicki (to whom the book is dedicated) should be proud of the result. It is not intended to be an exhaustive review of the field, but presents "some of the landmarks of past work in human heredity and some of the signposts for future development". The emphasis is on "the methodology . . . far more than the established facts" of human genetics. This is not, therefore, a book for beginning students in genetics or for medical students, but will be an invaluable aid to the graduate student and others actively working in human genetics.

After an introductory chapter describing the advantages, as well as the disadvantages of man as a subject for genetic study, the authors deal briefly but clearly with the physical basis of heredity. The inherited variations of the red blood cell are used as a text to present the concept of the genes, the specificity of their control of biochemical processes, and the immense number of different combinations in which they may occur. A chapter on "Nature and Nurture" emphasizes the complexity of gene-environment and gene-gene interactions and discusses the lines of evidence that may contribute information on the problem.

The chapters on dominant and recessive inheritance are original in considering the inheritance of rare and common genes separately, and including sex-linked and partially sex-linked dominant and recessive genes along with their autosomal counterparts, rather than treating them in a separate chapter. This section also discusses penetrance, pleiotropy, the Hardy-Weinberg law, the relation of age of onset of disease to more of inheritance, and the relation of consanguinity to recessive inheritance. Then there is a chapter on "Genes Neither Dominant Nor Recessive" which includes a good discussion of genetic "carriers" and a table summarising the diseases that may show a carrier state.

A discussion of quantitative inheritance includes a critical evaluation of the role of correlation in

estimating the contributions of heredity and environment to a given quantitative character, using height and intelligence as type examples. This is followed by a chapter on linkage in which the authors present Penrose's sib-pair methods of linkage detection (the Fisher-Finney method is considered too difficult, mathematically, for most readers) and a discussion of the unlikelihood that common genetic markers linked to pathological genes will be of any immediate practical use for genetic prognosis. In the next chapter the estimation of mutation rates and the problem of induced mutation in man are discussed clearly and critically.

After a section on physiological genetics dealing with a number of inherited metabolic defects in man, the reader reaches the meat of the book, "The Estimation of Genetic Parameters and Tests of Genetic Hypotheses". The use of the maximum likelihood method of estimation is demonstrated for a number of genetic situations (two and three autosomal alleles without and with dominance, sex-linked alleles, two pairs of alleles), and the use of χ^2 as a test of goodness of fit is illustrated. Here the mathematics becomes quite heavy, for which the authors make no apology, believing that "... a knowledge of certain branches of mathematics is no less essential to the serious student of human heredity than to the astronomer ...". The reviewer has not attempted to check the formulae or calculations. The following chapter deals efficiently with the knotty problem of ascertainment, though the difficulties involved are not dealt with as thoroughly as in a recent article of Schull's, which the authors have modestly omitted even from the bibliography. After some more advanced algebra the chapter concludes with a warning that it is no use applying fancy statistics to data that are inadequate, either through improper collection or insufficient understanding of the biological situation.

The chapter on population genetics deals with the frequencies of genes of universal distribution (e.g., blood groups), and of genes of restricted distribution (such as those for Thalassemia and the sickling phenomenon), and discusses the factors influencing these frequencies. Due note is taken of the possible errors and biases in estimating these factors, but the authors are optimistic about the potential contributions of research in this field to our understanding of anthropological problems. On the other hand, in the following chapter, they are pessimistic about the value of twin studies as a means of appraising nature-nurture interaction.

The final section of the book deals with the practical aspects of human genetics as applied in the fields of epidemiology, counselling, forensic medicine and eugenics. The authors have presented a cautious critical and constructive appraisal of the contributions of genetics in these fields, and this could be read with profit by doctors, social workers and others who have to deal with human families and their genetic problems, or who are otherwise concerned about the future of the human species. The book ends on a characteristic note of caution—the suggestion that "the effort which would be expended on a eugenics program might better go into efforts to explore the many gaps in our present fragmentary information."

F. Clarke Fraser
McGill University

The Unleashing of Evolutionary Thought

By Oscar Riddle. New York: Vantage Press. 1954, Pp. 414, $4.50.

This essay was born of a firm conviction that the attitudes and goals of society everywhere would be vastly matured by general acceptance of the concept of evolution. Since Dr. Riddle realizes that this conviction is shared by most scientists the book is primarily an explanation of why evolution is not more generally accepted. Nor is the book directed to scientists alone. It is written for "your neighbors and mine" who "are wholly unprepared to give thought to the things that would flow from a widely accepted view of the natural origin of man, of his biological and social nature, of the animal and social sources of morality, and of a world rid of the supernatural." For this reason Part I entitled "What Evolutionary Thought Is" has been added. The various chapters in this section cover such topics as, "The Problem of Creation", "Evolution and Ethics", "The Biological Inequality of Man". It may be seen that these are the everyday ideas which are most likely to be modified by an understanding of evolution in its broadest sense. For example, in the chapter entitled "Social Inheritance" Riddle lists three present day dangers. First, modern technological civilization fails to consider

sufficiently the biological exigencies of man. Second, organized religions hinder understanding of man's biological origin. Thus he is prevented from making sound plans for the future of society. Third, overpopulation brings into even sharper focus the need for application of eugenic measures.

It is curious that Dr. Riddle lists these dangers in this order for it has already become abundantly clear that the second is the main thesis of this book. Part II, the longest and most detailed, is called "Reins Held by Religion." Here is gathered an imposing mass of evidence to show that the organized religions hinder not only the dissemination of evolutionary ideas but also receptivity to these concepts. Riddle's most interesting contribution in Part II is his discussion of the manner in which evolution is taught in our secondary schools and universities. He is on firm ground here for he was chairman of a committee of the Union of American Biological Societies which studied the teaching of evolution in high schools of the United States. There is no gainsaying that the information obtained by this committee indicates that religious reasons, above all others, are at the basis of the gingerly way in which evolution is approached in our schools and colleges.

The third part "Opinion and Outlook" reaches strong, all-embracing conclusions. Dr. Riddle spares no religion here or abroad in fixing responsibility for the widespread ignorance of evolution and its implications.

Many, if not most, of the ideas herein have been published before. Many readers may think that the emphasis on religion is misplaced; that human nature resists evolutionary ideas because they are uncomfortable or disquieting for other reasons. Some will feel that the cogency of Dr. Riddle's argument is lost in a somewhat leaden style. These criticisms however, do not detract from the overall worthiness of his premises.

This book is not a pleasant one because Dr. Riddle does not spend much time being merely optimistic. He presents very little evidence that indicates any progress in the "pressing conflict" between belief in the supernatural and belief in what science can tell us about ourselves. Nevertheless, despite the mass of gloomy documentation one never entirely loses sight of Dr. Riddle's fundamental faith in man's "good purposes" and his "earnest doubts."

BEAL HYDE
University of Oklahoma